Psychometrika

COPYRIGHT, PSYCHOMETRIC CORPORATION, 1937

CONTENTS

VOLUME TWO SEPTEMBER 1937 NUMBER THREE

PSYCHOMETRIKA—VOL. 2, NO. 3
SEPTEMBER, 1937

SCALE VALUES DERIVED FROM THE METHOD OF CHOICES

J. P. GUILFORD
University of Nebraska

In many a psychometric problem, particularly in practical problems, the only data conveniently obtainable are the first choices given to certain stimuli among a list of stimuli that are available for selection. By assuming that such choices represent real comparative judgments, we may apply Thurstone's law of comparative judgments, extract experimental proportions from the numbers of first choices, and compute psychological scale values for the stimuli. Two procedures are proposed for estimating such proportions and examples of their applications are given. A procedure for allocating a meaningful zero point on the scale by the use of absolute judgments is explained and demonstrated. Suggestions are added for overcoming certain weaknesses and limitations of the method of choices.

One of the very earliest devices employed in evaluating aesthetic objects was Fechner's "method of choice" or *Wahlmethode*. From this parental stem have grown the methods of paired comparisons and of ranking and other variations. These derived methods have now been rationalized on the basis of probability theory and they have been made to yield quantitative measurements of stimuli on metric scales. The method of choice, in which each judge selects from among a series of stimuli laid out before him the one that to him seems greatest or best, is still lacking the rationale that would made it a useful measureing instrument.

Probably the method of choice, or to be more exact, the *method of first choices*, will never attain the level of respect with which some of the other methods are regarded, but its great practical usefulness and its wide applicability are strongly in favor of its continued use. Its utter simplicity of administration gives it a wide popular acceptance. The average individual who would be utterly bored if not irked by having to make numerous paired comparisons or even by having to place a small number of stimuli in rank order, very readily agrees to say which stimulus he likes best and he rarely has difficulty in doing so.

Some Applications of the Method of First Choices

A few examples will be cited to show where the method is used or can be used. In the typical election ballot, political or otherwise,

the voter makes a single choice among two or more candidates. When there are more than two, a simple paired comparison is no longer the type of judgment. When only one candidate is to be elected a simple plurality is obviously all that is required as a quantitative indicator. The relative vote-getting strength of the remaining candidates may be of interest to the candidates themselves and to their political supporters. The raw number of votes obtained by the candidates will serve as rough indicators of this, but this index may be misleading at times when taken merely at its face value. For example, a candidate who gets three times as many votes as a rival is not necessarily three times as popular or as strong as that rival. In nomination ballots when each voter writes in his sole first choice, and in run-off primary elections when a *number of candidates* are to be selected, the final list is not always ascertainable with validity and accuracy by a mere counting of votes. The differences between neighboring candidates in the ranked list, even though equal in terms of numbers of votes gained, do not give a correct picture of the spacing of the candidates along the continuum of public popularity or approval. One who wants a more exact, scientific picture of the outcomes of elections, in other words, must seek for a means of translating votes into scale positions on psychological continua.

The method of first choices is applied every day in hundreds of ways, whether tradesmen know it or not, in the great occupation of retailing. Customers who buy cars, radios, roofing, bathtubs, neckties, novels, and thousands of other products, in the public mart typically choose *one* at a time from among many samples. Not even an applied psychologist selects his clothing by a rigorous process of paired comparisons or of ranking or of rating. The alert retailer, wholesaler, and manufacturer, ever more aware of the problem of preferences of customers, do depend upon the volume of sales as an index of the popularity of their products. But even in the case of the more enlightened vendor there is as yet little conception of the development of psychological scales of satisfaction and the analysis of the product from the standpoint of its stimulating value in order to find the factors responsible for the scale value of the product. When there is a real awakening to the possibility of so analyzing and measuring customer reactions it will be found that "first choices" will typically furnish the raw data.

The two previous examples stress the practical use of the method of first choices. It also has its application to the solution of theoretical problems. As early as 1893, Jastrow made an attempt to evaluate

color preferences by asking thousands of individuals at the Chicago Columbian Exposition to record their first choices of colors and color combinations. Wherever large crowds congregate, as at expositions, museums, art galleries, and the like, an opportunity is afforded to collect data of aesthetic judgment from a relatively large population in a short time when only first choices are required. Had there been some ready device of extracting scale values from such data, no doubt much more desirable work of this sort would have been undertaken. There are other shortcomings of the method that have dissuaded investigators from its widespread use. The most serious of these defects is the fact that when there are a large number of stimuli from which to choose, some of them receive no choices at all. This gives us no definite idea of their relative worth. In addition to suggesting two processes for deriving scale values from the numbers of first choices, the writer will make other comments looking toward improvement in the method of choices.

The Treatment of Choices as Comparative Judgments

Let us state the underlying rationale of the method in as simple terms as possible. Let there be n stimuli from which one is to be selected as the best or greatest by each one of N individuals. Let C_a be the number of choices given to a particular stimulus R_a, and C_k the number of choices given to any other stimulus R_k. Let us assume that Thurstone's law of comparative judgment holds in the evaluation of stimuli in this judgmental situation.* The problem then is to estimate what proportion of the judges prefer R_a as against R_k, and to do this from the limited data that we get from first choices.

Now if 100 individuals make R_a their first choice, these 100 individuals preferred R_a to R_k. If only 40 individuals made R_k their first choice, nevertheless these 40 prefer R_k to R_a. From this information we have 140 individuals who have told us whether they prefer R_a or R_k. The proportion of those who prefer R_a to R_k is therefore given by the ratio

$$P_{a>k} = \frac{C_a}{C_a + C_k}. \tag{1}$$

Likewise,

$$P_{b>j} = \frac{C_b}{C_b + C_j},$$

*Thurstone, L. L., Psychological Analysis, *Amer. J. Psychol.*, 1927, **38**, 368-389.

and so on, every stimulus being paired with every other one until we have $n(n-1)$ proportions. From these are then determined the scale separations and finally the scale values as one does in the method of paired comparisons.*

There is one serious question, however, that arises immediately. This comes from the fact that the different proportions are not obtained from the same set of judges and not even from the same number of judges. The individuals who choose R_a and R_k are a totally different group than those who choose R_b and R_j, and so on for other pairs in which different stimuli are involved. If we know that the judges are all comparable, even roughly, so that one sub-group is replaceable by another without doing violence to the scale separations, then this difficulty is eliminated for practical purposes. The very large number of judges customarily used in this method helps to make the different sub-groups comparable. A satisfactory indicator of this condition should be found in the test of internal consistency. After the final scale values are computed, theoretical proportions can be deduced from them and if they closely resemble the observed proportions the question of systematic differences between sub-groups is well answered.† Evidence to be presented seems to show that sub-groups may be strikingly homogeneous with the total population of judges even when the number of judges is relatively small. As to the varying numbers of judges from which each proportion is obtained by the use of equation (1), this can be taken care of by a process of weighting the scale separations if one cares to do so.‡

The Computation of Scale Values

As a simple illustration of the procedure just explained, let us use some data obtained from the Pressey X—0 test. In this test there are many lists of five words each. In Part I the testee is instructed to encircle the one word in each list that he finds most unpleasant. In Table I below are given the first five words and the number of times they were encircled by 283 individuals. The range is from 5 for the word "aunt" to 113 for the word "suspicion." By the use of equation (1) the list of proportions given in Table I was computed. Table II gives the deviates, or scale separations, for pairs of stimulus words. The words have been arranged for convenience in the order of in-

*For a systematic treatment of paired comparisons, see the author's *Psychometric Methods*, New York: McGraw-Hill, 1936, Chapter VII.

†*Ibid.*, p. 230.

‡For a discussion of the weighting of scale separations, see L. L. Thurstone, The Measurement of Opinion, *Jour. Abn. and Soc. Psychol.*, 1928, **22**, 415-430.

creasing unpleasantness. At the bottom of Table II are given the sums and means of the columns. Since every difference between pairs is determinate and since none is greater than plus or minus 2.000, we may take the means of the columns to represent the true deviation of the scale values from the mean of them all, the unit being the standard error of the differences between stimuli on the psychological continuum of unpleasantness. Assuming Thurstone's Case V, in which the discriminal dispersions of all the stimuli are equal, we next multiply the means of the columns by $\sqrt{2}$ in order to obtain scale values in terms of the discriminal error of any one stimulus as the unit. The zero point is an arbitrary one and is at the mean of the scale values. Just how many of the words are actually below the indifference point on the affective scale we do not know from these data alone.

TABLE I

The Determination of the Proportions with Which One of Five Words Is Judged More Unpleasant Than Every Other One from First Choices Alone

	C_k	Proportions				
		1	2	3	4	5
1. aunt	5	.500	.868	.886	.949	.958
2. sex	33	.132	.500	.541	.738	.774
3. fear	39	.114	.459	.500	.705	.743
4. disgust	93	.051	.262	.285	.500	.549
5. suspicion	113	.042	.226	.257	.451	.500

TABLE II

Computation of the Scale Values from the Proportions Given in Table I

	1	2	3	4	5
1. aunt	.000	1.117	1.205	1.635	1.728
2. sex	—1.117	.000	.103	.637	.752
3. fear	—1.205	— .103	.000	.539	.653
4. disgust	—1.635	— .637	— .539	.000	.123
5. suspicion	—1.728	— .752	— .653	— .123	.000
Σ	—5.685	— .375	.116	2.688	3.256
M	—1.137	— .075	.023	.538	.651
$M\sqrt{2}$	—1.609	— .106	.033	.761	.921

In order to test the method further, five more lists of words from the Pressey X—0 test were selected at random, one more from Part I, two from Part III, and two from Part IV. In Part III the testee is instructed to encircle the word in each list that signifies an action or

trait that he considers most blameworthy. In Part IV he is told to encircle the thing about which he worries most. In Part III we may assume an ethical scale along which the practices are to be evaluated and in Part IV a "worry" scale. The words of the six lists are given in Table III, along with the number of times each one was encircled (C_k) and also the scale values as obtained by three different methods. The values under S_t were obtained by the use of formula (1) and Thurstone's Case V. The other two sets of scale values were obtained by procedures about to be explained. All scale values are given with reference to a real zero point which was determined in each case by procedures also to be described.

Before going into a description of these new procedures, a word should be said about the test of internal consistency and its results when applied to the scale values S_t. The sixty theoretical proportions were computed from the final scale separations. The discrepancies between these and the observed proportions were all very small. The largest was only .027, and this was obtained from an observed proportion that was derived from a ratio of 15/18, the most unreliable of all the proportions. Only three discrepancies were greater than .015, and three-fourths of them were less than .010. This very high agreement is probably due in part to the fact that the proportions are not completely independent. But aside from a certain amount of internal consistency imposed upon the scale separations by the use of first choices, as is also true when rankings are used, the exceptionally close agreement between expected and obtained proportions indicates that although some proportions arise from mutually exclusive sub-groups of judges, either the sub-groups are highly comparable, or else what errors are introduced because of their incomparability somehow cancel out in the computation of scale values. The agreement also lends support to the assumption of Thurstone's Case V.

A Shorter Method Assuming a Composite Standard

The second procedure for deriving scale values S_c is shorter in operation than the first. It assumes a single composite standard or level, composed of all the stimuli, including the stimulus R_a whose probability of preference we seek to determine.* Let C_a again stand for the number of first choices of R_a. But C_a is conceived as the number of choices over each and every other stimulus taken separately.

*A similar assumption has been made for dealing with paired comparisons. See author's *op. cit.*, p. 236.

With R_b as the competitor, out of $C_a + C_b$ assumed comparisons, R_a receives C_a choices; with stimulus R_c as the competitor, out of $C_a + C_c$ comparisons, R_a receives also C_a selections; and so on, for all other stimuli in the list. Let us suppose that the same reasoning holds for the comparison of R_a with itself, since R_a makes up a part of the composite standard. To continue the same line of reasoning, when R_a is compared with itself there are $2C_a$ comparisons, out of which there are C_a choices, as usual. The proportion when R_a is compared with itself is $C_a/2C_a$ *or* .500.

To pool all the judgments of R_a so obtained, we have R_a chosen a total of nC_a times. The total number of times that R_a appears for comparison is the sum of the combinations of first choices of all the

TABLE III

Thirty Words from the Pressey *X*—0 Test and Their Scale Values

	Number of Choices		Scale Values				Number of Choices		Scale Values		
	C_k	c_k	S_t	S_c	S_i		C_k	c_k	S_t	S_c	S_i
aunt	5	12	—2.00	—1.93	—1.73	sidewalk	3	5	—2.30	—2.52	—1.64
sex	33	62	— .50	— .43	— .78	roar	15	60	—1.08	—1.34	— .80
fear	39	115	— .36	— .28	— .24	dislike	37	110	— .34	— .57	— .26
disgust	93	196	+ .37	+ .49	+ .50	wiggle	96	151	+ .43	+ .27	+ .33
suspicion	113	193	+ .53	+ .66	+ .47	divorce	130	183	+ .67	+ .53	+ .52
smoking	14	97	— .65	— .70	— .40	innocence	3	21	—2.60	—2.68	—1.45
flirting	19	102	— .40	— .44	— .35	meekness	28	69	— .88	— .96	— .69
spitting	69	186	+ .70	+ .68	+ .41	dullness	36	106	— .68	— .75	— .32
begging	89	199	+ .91	+ .90	+ .54	weakness	84	131	+ .02	.00	— .09
swearing	91	221	+ .93	+ .92	+ .79	ignorance	131	177	+ .38	+ .38	+ .33
inventions	1	5	—3.15	—3.35	—2.10	cats	14	32	—2.10	—2.05	—1.21
awkwardness	63	143	— .13	— .25	+ .02	teacher	52	90	— .98	— .92	— .47
death	69	125	— .05	— .17	— .14	engagement	59	75	— .88	— .81	— .62
wreck	70	131	— .04	— .15	— .09	epilepsy	64	86	— .79	— .74	— .51
insanity	79	121	+ .06	— .05	— .18	confusion	93	121	— .48	— .42	— .18

stimuli taken two at a time. In symbolic terms it is $\Sigma(C_a + C_k)$, in which C_k stands for the number of choices for each stimulus in turn, including R_a. The proportion is given by the ratio

$$p_{a>cs} = \frac{nC_a}{\Sigma(C_a + C_k)} \qquad (2)$$

in which $p_{a>cs}$ is the probability that stimulus R_a is judged greater than the composite standard, and the other constants are as defined in the preceding discussion.

Equation (2) can be simplified for practical purposes as follows. The expression $\Sigma(C_a + C_k)$ can be written as $nC_a + \Sigma C_k$. Since $\Sigma C_k = N$, the equation reduces to

$$P_{a>cs} = \frac{nC_a}{nC_a + N}. \tag{3}$$

Dividing through by n,

$$p_{a>cs} = \frac{C_a}{C_a + \frac{N}{n}}. \tag{4}$$

The second procedure, using formula (4), may be very briefly illustrated by means of the same data as are given in Table I. The solution of scale values through this procedure is shown in Table IV.

TABLE IV

Solution of Scale Values for the Unpleasantness of Words from the Numbers of First Choices, Assuming a Composite Standard

	C_k	$C_k + \frac{N}{n}$	$p_{a>cs}$	X_{cs}	$\sqrt{2}X_{cs}$	S_t
1. aunt	5	61.6	.042	—1.398	—1.977	—1.609
2. sex	33	89.6	.219	— .337	— .477	— .106
3. fear	39	95.6	.406	— .233	— .330	+ .033
4. disgust	93	149.6	.681	+ .311	+ .440	+ .761
5. suspicion	113	169.6	.693	+ .432	+ .611	+ .921

One uncertainty in this method is the unit of the scale. It was found empirically that in the six sets of data used here for illustrative purposes, the range and standard deviation of the scale values from the Thurstone method (S_t) were about forty per cent greater than those found by the composite standard method. To make the values from the two procedures more comparable, the scale values from the latter method were multiplied by $\sqrt{2}$ just as they were in Thurstone's method. The last two columns of Table IV show the comparison of the two sets of values. While they are now approximately equivalent in range, the zero points do not coincide. The zero point for the S_t values is their mean while the zero point for the other set of values is .343 above their mean. In neither case is the zero point located at a meaningful point among the list of words. The problem of determining a meaningful zero point is the one to which we turn next.

The Location of a Meaningful Zero Point

A zero point with psychological meaning cannot be determined from paired comparisons, rankings, or first choices alone. The writer has previously suggested a device for locating the zero point when it is conceived as an indifference point associated with scales such as that for affective value.* For the words selected from Part I of Pressey's test, for example, the natural division point is between those words that are judged as unpleasant less than 50 per cent of the time and those judged unpleasant at least 50 per cent of the time. In Part III the division point comes at those practices or traits that are judged ethically wrong 50 per cent of the time. In Part IV a zero point may be conceived at those items about which 50 per cent of the individuals confess their worries.

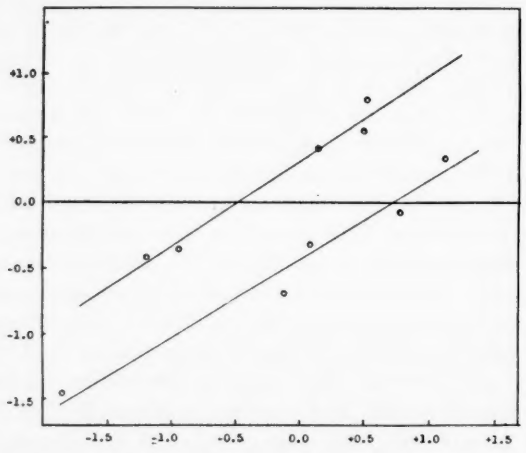
From the responses to the words in the Pressey test we fortunately have the necessary data for making use of the definitions of the zero points just presented. Each testee was instructed not only to encircle the worst word in each list of five, but also to cross out all words that were to him either unpleasant, morally wrong, or about which he worried. The responses of the individuals were thus forced into two categories by this second instruction. From the proportion of the reactions in either category, crossed or uncrossed, for every word, we can determine the scale separation of the word from the zero point of the scale. The scale separation is conceived as a standard measure or deviate. The procedure is so relatively simple that it will not be demonstrated here. Table III contains the numbers of times all words were crossed out, under the symbol c_k, and the scale values derived from them, under the symbol S_i. All the words having negative S_i values in Table III may be regarded as "not unpleasant," "not wrong," and "not worriers" for the population of students who reacted to them. It would not be correct without further information to say that those words with negative values are necessarily "pleasant," "right," or "non-worriers," however. Had the testees been asked to react with either "pleasant" or "unpleasant" for the first set of words, "right" or "wrong" for the second set, and "worrier" or "non-worrier" for the third set, the positions of the zero points might have been different.

If there is an actual indifference *point*, it is at the median of a *zone* of indifference. In order to locate this ideal point exactly, a different set of categories than that employed for any class of words in

Op. cit., p. 239.

the Pressey test would have to be used. Were a third category of "indifferent" permitted, doubtless a large number of such judgments would have been given. These occupy a large part of the zone of indifference. Judgments of the type imposed by the Pressey test lend themselves only to establishing a division point between the "unpleasant" on the one hand, and the "indifferent" and "pleasant" combined on the other. There would conceivably be a similar division point between the "indifferent" and the "pleasant." The ideal indifference point lies midway between these two points.

Defining the zero points of our three scales in question as they are defined above, however, we may proceed to allocate similar zero points for the scale values S_t and S_c. Since the range of values for the data from the absolute judgments (cross-outs) is obviously smaller than that for the other two methods, we cannot simply compare means and then add a constant amount where it is necessary to do so.



SCALE VALUES DERIVED FROM FIRST CHOICES

Fig. 1. An illustration of the graphic solution of the zero point on a scale derived from comparative judgments.
(Caption for the ordinate: Scale Values Derived from Absolute Judgments).

The functional relationship between the S_i values and the others is clearly linear when the points are plotted. Least square methods could be employed to solve the problem, to find what values on the S_t and S_c scales are equivalent to zero on the S_i scale. For the sake

of a quick solution we may employ a graphic method as is illustrated in Fig. 1. Here the S_i values are plotted against the S_t values for two typical sets of words. One is the set including the words "smoking, flirting, spitting, begging, and swearing," and the other contains the words "innocence, meekness, dullness, weakness, and ignorance." The position of the zero point on the S_t scale in each case is that value at which the line of best fit crosses the level of zero on the S_i scale. Those points are —0.50 and +0.75 for the two sets of words, respectively.

One could, of course, be content with the S_i scale alone. It carries with it its own definite and meaningful zero point. Its unit, however, as was said before, is an unknown quantity. Since its scale values bear a linear relationship to those of the S_t scale, the necessary correction can be made in the unit when the S_t values are known. Between the two methods, the S_t scale furnishing the unit and the S_i scale furnishing the zero point, we have the makings of a complete psychological yardstick, and it should often be advantageous to employ them together. One advantage of the S_i scale over the other two that is apparent in the illustrative problems, is that direct comparisons can be made between words appearing in different lists of stimuli when the type of judgment remains the same. In the other two procedures, unless the position of zero is established somehow, scale values may be compared only within their own groups.

It is of interest, in Table III, to compare the values in scale S_i with those in the other two, since they are derived from different types of judgments. The former were derived from "absolute" judgments whereas the latter came from comparative, or at least "pseudo-comparative" judgments. The correlations between the S_t and S_c scales are well nigh perfect, since they come from the same data and very similar treatment of those data. The S_i values show some discrepancies from the other two, the correlations being in the neighborhood of .90. The testee may cross out words in a list of five that he does not encircle; but he also encircles words that he does not cross out, at times. He may find no words in the five that are unpleasant, wrong, or worrisome and yet by instruction he must encircle the one tending most in one of those directions. No glaringly significant discrepancies can be seen in Table III, even in algebraic sign. But it is possible for such discrepancies to occur and when they do an analysis of the reasons is in order. This is another reason for using the two types of judgments, relative and absolute, in the scaling of mental values.

The Method of Last Choices

There still remains something to be said concerning a supposedly fatal weakness in the method of first choices. This is the fact that some stimuli in a series may never receive any first choices, thus leaving their scale values indeterminate. The obvious solution, extending the principle of first choices, is to ask the judges for *last* choices as well. It would be reasonable to apply the same rationale and procedures to last choices as have been recommended for first choices. It may be that the last choices for some judges are the first choices for others. The scales derived from both ends, in this event, would serve as valuable checks upon each other. Asking a judge for last choice as well as first increases his labor only slightly but it adds much needed information and verification for the use of the experimenter. Probably no stimulus will be left without one kind of choice or the other if this double task is assigned to the judges, and no stimulus will be left with an indeterminate value. From the standpoint of the judge, the use of last choices as well as first would tend to stabilize his conception of the continuum along which he is to pass judgment. With this addition, then, and with the prospect of deriving rational scale values, the method of choices should once more take its place among its more favored progeny and again come into serious use.

PSYCHOMETRIKA—VOL. 2, NO. 3
SEPTEMBER, 1937

THE THEORY OF THE ESTIMATION OF TEST RELIABILITY

G. F. KUDER AND M. W. RICHARDSON
The University of Chicago

The theoretically best estimate of the reliability coefficient is stated in terms of a precise definition of the equivalence of two forms of a test. Various approximations to this theoretical formula are derived, with reference to several degrees of completeness of information about the test and to special assumptions. The familiar Spearman-Brown Formula is shown to be a special case of the general formulation of the problem of reliability. Reliability coefficients computed in various ways are presented for comparative purposes.

The reliability coefficient is of interest because it gives, by the simple assumption that a test score has two components, viz., true score and variable error, an (indirect) estimate of the random error variance present in an obtained test score variance. No matter how computed, the reliability coefficient is only an *estimate* of the percentage of the total variance that may be described as true variance, i.e., not due to error.

The usual methods of estimating test reliability are too well known to justify description here. These methods differ in such a fashion that no close estimate can be made of the results of one method, knowing the estimate obtained by another method. It is always desirable, even necessary, for the investigator to state how he made his estimate of the reliability coefficient.* The retest coefficient on the same form gives, in general, estimates that are too high, because of material remembered on the second application of the test. This memory factor cannot be eliminated by increasing the length of time between the two applications, because of variable growth in the function tested within the population of individuals. These difficulties are so serious that the method is rarely used.

Although the authors have made no actual count, it seems safe to say that most test technicians use the split-half method of estimating reliability. This method involves an arbitrary division of the test

*The critical reader will reflect that, in addition, the investigator must report the range, or better, the variance of the group tested. The present study is not concerned with that matter.

into two parts, and the computation of the correlation-coefficient of the two sets of scores thus derived. The correlation coefficient thus obtained is taken as an estimate of the reliability of either half, and the Spearman-Brown formula for double length is then used to estimate the reliability coefficient of the whole test. The split-half method is commonly supposed to give estimates that are too high; this is an uncertain generalization unless one has some definitely defensible standard. A more pertinent observation about the split-half coefficient is that *it is not a unique value.* There are $\frac{n!}{2(\frac{n}{2}!)^2}$ different ways of dividing a test of n items into two halves. Each one of these ways of splitting the test gives its own estimate of the reliability coefficient.* True enough, not all these ways of splitting are equally defensible on *a priori* grounds. It remains true, however, that there are large fluctuations in the value of the reliability coefficient as obtained from diffdiferent ways of constituting the two halves.†

The supposedly best method of estimating the reliability coefficient is to find the correlation between two *equivalent* forms, given at the same time. The crux of the matter here is *equivalence.* Actually the difficulties discussed in connection with the split-half coefficient still apply, in perhaps smaller degree. Again, there is no unique value of the reliability coefficient. In the quest for equivalence, the shift of items from one form to the other will affect the magnitude of the coefficient. In this situation, there are $\frac{(2n!)}{2(n!)^2}$ different coefficients, again not equally defensible.

In view of the limitations briefly described in the foregoing, the authors present certain deductions from test theory which lead to unique values of the reliability coefficient.‡ The least exact approximation we shall describe involves assumptions no more unreasonable

*With certain assumptions as to the distribution of inter-item correlations it would be possible to estimate, theoretically, the expected distribution of reliability coefficients thus to be computed. The most representative value (perhaps the mean) could then be taken as the best estimate and the problem thus solved. It is likely, however, that the solution would be enormously complicated by the possibility that the matrix of inter-item coefficients would have a rank greater than one. See Mosier, Charles I., "A Note on Item Analysis and the Criterion of Internal Consistency," *Psychometrika*, 1936, **1**, pp. 275-282.

†Brownell Wm. A., "On the Accuracy with which Reliability May Be Measured by Correlating Test Halves," *J. Exper. Educ.*, 1933, 1, pp. 204-215.

‡It should be mentioned that the main outlines of the simple argument in this article were derived independently by the two authors. In a chance conversation it developed that the two had reached similar conclusions by methods similar in principle.

than those basic to the Spearman-Brown formula. Any one of the formulas will give a unique estimate of the coefficient in all situations to which it is applicable. In certain cases, the commonly calculated parameters of the test score distribution will afford, in two minutes of time, a fairly good estimate of the reliability coefficient.

We shall consider a test variable t made up of n unit-weighted items applied to a population of N individuals. In the general case, we shall allow for the possibility of the inter-item coefficients varying between their possible limits, and also for varying proportions of correct answers; items need not be equally difficult or equally correlated with other items. This enables us to state the formally complete and theoretically most exact method of estimating the reliability of test t.

CASE I.

The data required are the number of items in the test, the difficulties of the items, the inter-item correlations, and the standard deviation of the total test. In one of the possible solutions suggested it is assumed that the matrix of inter-item correlations has a rank of one.

The correlation between two forms of a test is given by

$$r(a+b+\cdots+n)(A+B+\cdots+N) = \tag{1}$$

$$\frac{r_{aA}\,\sigma_a\,\sigma_A + r_{aB}\,\sigma_a\,\sigma_B + \cdots + r_{n(N-1)}\,\sigma_n\,\sigma_{N-1} + r_{nN}\,\sigma_n\,\sigma_N}{[\sigma_a^2+\sigma_b^2+\cdots+\sigma_n^2+2(r_{ab}\,\sigma_a\,\sigma_b + r_{ac}\,\sigma_a\,\sigma_c+\cdots+r_{n(n-1)}\,\sigma_n\,\sigma_{n-1})]^{\frac{1}{2}}\times [\sigma_A^2+\sigma_B^2+\cdots\sigma_N^2+2(r_{AB}\,\sigma_A\,\sigma_B + r_{AC}\,\sigma_A\,\sigma_C+\cdots+r_{N(N-1)}\sigma_N\,\sigma_{N-1})]^{\frac{1}{2}}},$$

in which a, b, $\cdots$ n are items of the test, and A, B, $\cdots$ N are corresponding items in a second hypothetical test. Equivalence is now defined as interchangeability of items a and A, b and B, etc.; the members of each pair have the same difficulty and are correlated to the extent of their respective reliabilities. The inter-item correlations of one test are the same as those in the other. These relationships constitute the operational definition of *equivalence* which is to be used.*

By this definition of equivalence, the two expressions in the denominator of equation (1) are identical. It may then be seen that the numerator and denominator are the sums of the entries in square tables which are the same except for the entries in the principal diagonals. The entries in the principal diagonal of the numerator are the reliabilities of the items multiplied by their variance, while the entries in the diagonal of the denominator are merely the variances of the items. The formula for test reliability then becomes:

*It should be noted that this definition of equivalence is more rigid than the one usually stated.

$$r_{tt} = \frac{r_{aa}\,\sigma_a^2 + r_{bb}\,\sigma_b^2 + \cdots + r_{nn}\,\sigma_n^2 + 2(r_{ab}\,\sigma_a\,\sigma_b + r_{ac}\,\sigma_a\,\sigma_c + \cdots + r_{n(n-1)}\,\sigma_n\,\sigma_{n-1})}{\sigma_a^2 + \sigma_b^2 + \cdots + \sigma_n^2 + 2(r_{ab}\,\sigma_a\,\sigma_b + r_{ac}\,\sigma_a\,\sigma_c + \cdots + r_{n(n-1)}\,\sigma_n\,\sigma_{n-1})}. \qquad (2)$$

The denominator of equation (2) is simply the expression for the variance of the sum of the items a to n, when each item is given a score of one. We can therefore substitute σ_t^2, the obtained variance of test scores, directly in the denominator, and also in the numerator by use of a suitable correction.

In order to write the numerator term, we must adjust the variance for the fact that the entries in the diagonals of the numerator and denominator tables are different. We therefore subtract from the obtained variance the sum of the variances of the items ($\sum_1^n pq$) and substitute the sum of the products of the variance and reliability of each item ($\sum_1^n r_{ii}pq$). The variance of any item i is p_iq_i.

The formula then becomes

$$r_{tt} = \frac{\sigma_t^2 - \sum_1^n pq + \sum_1^n r_{ii}pq}{\sigma_t^2}, \qquad (3)$$

where σ_t^2 is the obtained test variance, $\sum_1^n pq$ is the sum of item variances, and $\sum_1^n r_{ii}pq$ is the sum of the products of item reliabilities and their variances.

Equation (3), while basic, is not adapted to calculations, because the r_{ii}'s are not operationally determinable except by use of certain assumptions. However, certain approximations are possible. If the inter-item correlations are available, two methods of estimating the n different values of r_{ii} suggest themselves. One is to use the average correlation of item i with the n—1 other items of the test as an estimate of the reliability of item i. This method, or other methods, of estimating the reliability of an item may be thought to be crude; however, it will be noted by reference to the square tables previously suggested that the r_{ii}'s comprise for a 100-item test only one per cent of the total number of entries whose values enter into the determination of the reliability coefficient of the whole test. Reasonable guesses as

to the values of r_{ii} would probaly not affect the final result very much, unless the tests were very short.

Another method is to estimate the unknown r_{ii} as the average computed from all the second-order minors of the matrix of inter-item correlations in which r_{ii} is the single unknown. By this method,

$$r_{ii} = \frac{\Sigma \frac{r_{ij}\, r_{ik}}{rjk}}{\frac{1}{2}(n-1)(n-2)}, \tag{4}$$

where i, j, and k are all different, and where the Σ means the sum of the separate determinations of r_{ii} from the $\frac{1}{2}(n-1)(n-2)$ minors. This method assumes that the matrix is of rank one, or that the test measures one function. This method would be justified only where n is fairly small.

CASE II.

The data required are the numbers of items in the test, the difficulties of the items, the item-test correlations, and the standard deviation of the test. It is assumed that the matrix of inter-item correlations has a rank of one.

A more usable approximation is adapted to those situations in which an item analysis giving values of item-test correlations has been made. If we care to assume that item and test measure the same thing (which, of course, we do when we put the item into the test), we may write

$$\frac{r_{it}}{\sqrt{r_{ii}\, r_{tt}}} = 1, \tag{5}$$

where r_{it} is the correlation between the item and the test, r_{ii} and r_{tt} are the reliabilities of item and test respectively.

Then

$$r_{ii} = \frac{r_{it}^2}{r_{tt}}. \tag{6}$$

Substituting $\frac{r_{it}^2}{r_{tt}}$ for r_{ii} in equation (3), we have

$$r_{tt} = \frac{\sigma_t^2 - \sum_1^n pq + \frac{\sum_1^n r_{it}^2 pq}{r_{tt}}}{\sigma_t^2}. \tag{7}$$

Solving for r_{tt}:

$$r_{tt} = \frac{\sigma_t^2 - \Sigma pq}{2\sigma_t^2} \pm \sqrt{\frac{\Sigma r_{it}^2 pq}{\sigma_t^2} + \left(\frac{\sigma_t^2 - \Sigma pq}{2\sigma_t^2}\right)^2}. \tag{8}$$

In practice, only the positive value of the radical in the right member of the equation is admissible. Equation (8) gives an estimate of the reliability coefficient in those situations in which the techniques of item analysis have been applied. In each case, Σ denotes summation over the n items.

CASE III.

The data required are the number of items in the test, the difficulties of the items, and the standard deviation of the test. It is assumed that the matrix of inter-item correlations has a rank of one and that all intercorrelations are equal.

In other situations, we may be willing to assume that the items are equally intercorrelated, but allow their difficulties to vary over a wide range. We shall proceed, therefore, to investigate this case. By assuming r_{ij} to be constant and equal to $\bar{r}_{ii}$ in equation (2) we have

$$r_{tt} = \frac{\bar{r}_{ii}\left(\sum_{i=1}^{n} \sqrt{p_i q_i}\right)^2}{\sigma_t^2}, \tag{9}$$

in which $\sqrt{p_i q_i}$ is the standard deviation of item i. Equation (9) gives an estimate of the reliability coefficient. An approximation to equation (9) is given by

$$r_{tt} = \frac{\bar{r}_{it}\Sigma\sqrt{pq}}{\sigma_t} \tag{10}$$

by assuming $\bar{r}_{ii} = \frac{\bar{r}_{it}^2}{r_{tt}}$, where $\bar{r}_{it}$ is the average item-test coefficient.

Since the test t is the sum of its items $a, b, \cdots n$, the variance of test scores is given by

$$\sigma_t^2 = \sigma_a^2 + \sigma_b^2 \cdots + \sigma_n^2 + 2(r_{ab}\sigma_a\sigma_b + r_{ac}\sigma_a\sigma_c + \cdots + r_{(n-1)n}\sigma_{n-1}\sigma_n, \tag{11}$$

in which $a, b, \cdots n$ are items of the test.

If all intercorrelations are assumed equal ($\bar{r}_{ii}$), and $\sqrt{p_i q_i}$ is used as the σ for an item,

$$\sigma_t^2 = (\Sigma\sqrt{pq})^2\bar{r}_{ii} - \Sigma pq\bar{r}_{ii} + \Sigma pq, \tag{12}$$

in which

$$\Sigma\sqrt{\overline{pq}} = \text{sum of the } \sqrt{\overline{pq}}\text{'s for items } a \text{ to } n,$$

and

$$\bar{r}_{ii} = \frac{\sigma_t^2 - \Sigma pq}{(\Sigma\sqrt{\overline{pq}})^2 - \Sigma pq}. \tag{13}$$

Substituting for $\bar{r}_{ii}$ in formula (9)

$$r_{tt} = \frac{\sigma_t^2 - \Sigma pq}{(\Sigma\sqrt{\overline{pq}})^2 - \Sigma pq} \cdot \frac{(\Sigma\sqrt{\overline{pq}})^2}{\sigma_t^2}. \tag{14}$$

Again, all summations are over the items.

This formula is recommended for use when there is reason to believe that the inter-item correlations are approximately equal.

We shall digress slightly to illustrate the degree of approximation involved in the various steps. Let us suppose that, with reference to equation (9), we are in addition willing to assume equal standard deviations of items. With such an assumption, we have

$$r_{tt} = \frac{r_{ii} n^2 \overline{pq}}{\sigma_t^2}. \tag{15}$$

in which $\overline{pq}$ is the average item variance.

But

$$\sigma_t^2 = n\overline{pq}\,[1 + (n-1)\,r_{ii}], \text{ from (12)} \tag{16}$$

by similar assumptions. Substituting (16) in (15), we have

$$r_{tt} = \frac{n\bar{r}_{ii}}{1 + (n-1)\bar{r}_{ii}}. \tag{17}$$

Equation (17) is, of course, the familiar Spearman-Brown formula, which is predicated upon test length as the only variable affecting reliability, given a constant value of the reliability of the element.

It is now convenient to introduce another variant of equation (3), with assumptions similar to those involved in the Spearman-Brown formula.

From equation (12),

$$r_{ii} = \frac{\sigma_t^2 - n\overline{pq}}{(n-1)\,n\overline{pq}}, \tag{18}$$

since $\Sigma pq = n\overline{pq}$.

Substituting this value of r_{ii} in equation (15) we have

$$r_{tt} = \frac{\sigma_t^2 - n\overline{pq}}{(n-1)n\overline{pq}} \cdot \frac{n^2\overline{pq}}{\sigma_t^2}, \tag{19}$$

which simplifies to

$$r_{tt} = \frac{n}{n-1} \cdot \frac{\sigma_t^2 - n\overline{pq}}{\sigma_t^2}. \tag{20}$$

Equation (20) gives an estimate of the reliability of a test, knowing the number of items, the standard deviation, and the *average* variance of the items. This would not seem to be ordinarily a useful formula since it requires essentially the same basic data as formula (14), but involves one more approximation. Empirical evidence presented at the end of this paper, however, shows that reliabilities obtained by formula (20) do not for the tests used, vary more than .001 from those obtained from formula (14). Since formula (20) eliminates the necessity for computing $\sqrt{\overline{pq}}$ for each item, it accomplishes a material saving in labor. It serves, too, as a basis for the formula recommended for use in Case IV.

CASE IV.

The data required are the number of items in the test and the standard deviation and mean of the total scores. It is assumed in this case that the matrix of inter-item correlations has a rank of one, that these correlations are equal, and that all items have the same difficulty.

Solution of formula (20) becomes greatly simplified if we make the rigid assumption that all items have the same difficulty. As the formula now stands it is necessary to obtain the average variance. The average variance ($\overline{pq}$) is equal to the product of average p and average q, ($\bar{p}\,\bar{q}$), if the items all have the same difficulty. In this case,

$$r_{tt} = \frac{n}{n-1} \cdot \frac{\sigma_t^2 - n\bar{p}\,\bar{q}}{\sigma_t^2}. \tag{21}$$

The average value of p may be easily obtained from the formula

$$\bar{p} = \frac{\Sigma X_t}{nN} = \frac{M_t}{n}, \tag{22}$$

when ΣX_t is the sum of the scores of N subjects on a test of n items, and M_t is the mean of the test scores.

The difference between equations (20) and (21) should be noted.

Equation (20) calls for the average of the item variances ($\overline{pq}$); equation (21) calls for the average of the item difficulties ($\bar{p}$) and this value subtracted from 1.00, ($\bar{q}$). When all items have the same difficulty, $\overline{pq}$ is equal to $\bar{p}\,\bar{q}$, but if there is variation in difficulty among the items, $\bar{p}\,\bar{q}$ becomes larger than $\overline{pq}$, and this discrepancy increases as the variation increases. This means that the estimate of reliability obtained by formula (21) is equal to or less than that obtained by formula (20). If Equation (22) is used to get an estimate of $\bar{p}$, the reliability coefficient can be quickly estimated from the mean, standard deviation, and the number of items. This formula may be regarded as a sort of foot-rule method of estimating test reliability without the necessity of splitting halves, rescoring twice, and calculating a correlation coefficient. According to theory and to the applications already made, the formula may be expected to give an underestimate of the reliability coefficient in situations not favorable for its application. If Equation (21) should give a higher value than the split-half, one would suspect the latter of being abnormally low because of some unfavorable way of splitting. The split-half Spearman-Brown coefficient cannot be regarded as the standard from which to judge other estimates. The split-half method involving use of the Spearman-Brown formula may produce estimates of reliability which are either too high or too low. Reliabilities obtained from the formulas presented here are never overestimates. When the assumptions are rigidly fulfilled, the figures obtained are the exact values of test reliability as herein defined; if the assumptions are not met, the figures obtained are underestimates.

It may be useful to suggest an interpretation of Equation (21) which has some bearing on the general problem of reliability. For r_{tt} to be positive, σ_t^2 must exceed $n\bar{p}\bar{q}$. Now $n\bar{p}\bar{q}$ is the variance of n equally difficult items when they are uncorrelated, by the familiar binomial theory.* Hence r_{tt} is positive for any average inter-item correlation that is positive. But negative reliability is inadmissible; hence only to the extent to which test items are positively intercorrelated will a test have reliability. It is implicit in all formulations of the reliability problem that *reliability is the characteristic of a test possessed by virtue of the positive intercorrelations of the items composing it.*

Table I presents a comparison of reliability coefficients computed

*Dunlap, J. W. and Kurtz, A. K., *Handbook of Statistical Nomographs, Tables and Formulas*, World Book Company, New York. Formula No. 46.

by equation (21) with a split half coefficient for various tests. The time of computation was approximately two minutes for each test, applying Equation (21).

TABLE I

Test No.	Nature	Range of values of p	$\bar{p}$	n	σ_t	Reliability Coefficient	
						By equation (21)	Split-half, Spearman-Brown
1	College Achievement	.05-.22	.156	50	6.56	.864	.880
2	"	.23-.40	.318	50	9.24	.891	.906
3	"	.41-.59	.522	50	10.96	.914	.923
4	"	.60-.77	.672	50	8.69	.872	.896
5	"	.78-.95	.852	50	6.57	.871	.888

Table II presents results from several formulas. As in Table I, three decimal figures are retained, merely to illustrate the differences obtained by the various formulas.

TABLE II

Test No.	Nature	Mean Score	n	σ_t	Reliability Coefficient, as estimated by			
					Case II Equation (8)	Case III Equation (14)	Case III Equation (20)	Case IV Equation (21)
6	multiple choice vocabulary	24.39	65	7.62	.823	.808	.808	.733
7	do	24.13	65	7.92	.839	.826	.825	.758
8	general information	.....	25	.....	.729	.716	.716	.714

The foregoing results are not intended to confirm the theory developed, but they may serve to illustrate the degree of divergences of results that may be expected in actual application. In comparing these estimates, it should be noted that all the tests are short; longer tests may be expected to give less variable estimates. Several algebraic variants are not here presented; they may be easily derived when their use is indicated. The choice of formula to be used in any actual situation will depend upon the amount of information about the components of the test, and upon the degree of accuracy desired. It is the belief of the authors that in many cases the quick estimate afforded by Formula (21) may be good enough for all practical purposes; if the items vary greatly in difficulty, Formula (20) appears to be adequate in any case.

PSYCHOMETRIKA—VOL. 2, NO. 3
SEPTEMBER, 1937

STUDIES IN THE LEARNING FUNCTION
III. An Interpretation of the Semi-Major Axis*

L. E. WILEY AND A. M. WILEY
Ohio Wesleyan University, Delaware, Ohio

An interpretation of the semi-major axis of the Thurstone-type learning curve fitted by the authors to animal maze data is given. This interpretation affords a quantitative description of "insight".

In two recent publications we have shown that the theoretical learning curve developed by L. L. Thurstone fits learning data collected from animals with cerebral cortical injury which were trained on the Lashley-type maze.

Thurstone's curve is of the form

$$u = \frac{\sqrt{m}}{aK} - \frac{\sqrt{m}}{K} \cdot \frac{u}{R}, \qquad (1)$$

in which u represents the accumulated errors, R the number of trials, m the difficulty of the problem, k the learning constant of the animal, and a is an arbitrary constant that can be absorbed in either m or k. The coordinates are measured from the theoretical point at which learning begins.

Thurstone has discussed the relation of the constants of the curve to insight. He states,

> "One of the current problems in learning concerns the interpretation of insight. Here we may look upon a rather sudden rise to perfection as the objective and quantitative evidence of insight although there are other qualitative and really more convincing forms of evidence for it. The sudden rise in the learning curve to the level of perfection may be expected by our fundamental equation under either of two conditions. These conditions are (a) a value of k considerably higher than unity and (b) a low value of m. Strangely enough, the phenomenon of insight may be found, according to our equation, at the two opposite ends of the scale of learning, namely, (a) in highly rational or conceptual learn-

*A grant-in-aid from the National Research Council has made possible the series of investigations, of which this paper is a part. We wish to thank Dr. L. L. Thurstone for criticisms and suggestions.

ing (k is large) and (b) in the simplest possible forms of animal trial-and-error learning (m is low)."

Whether or not insight is present depends upon the solution for m and k. Since this is not possible unless we have more than one learning situation, we have set up another aspect of the learning curve which quantifies and describes the limits of learning.

Since in experimental work we never do know where learning begins we have changed the coordinates of the system to the point where training begins. The curve then takes the form

$$w' = A + \frac{B R'}{C + R'}, \tag{2}$$

in which w' represents the accumulated errors, R' is the number of trials, A, B, and C are constants. The relationship between the two systems of coordinates is expressed by

$$w' = u + e \tag{2}$$

$$R' = R + r$$

in which e represents the number of errors the rat makes before he begins to learn and r represents the number of trials he runs before he begins to learn.

Figure 1 represents the learning curve in its relationship to the theoretical asymptotes. OR'_a is the horizontal asymptote of the curve, the upper limit of learning. As the number of trials is extended to infinity, the number of errors approaches the limit set by OR'_a. Ow'_a is the vertical asymptote of the curve. As the number of errors approaches negative infinity, the number of trials theoretically approaches the limit set by the line Ow'_a. O is the point at which the two asmyptotes intersect.

In the first publications we have pointed out the fact that the learning curve is an equilateral hyperbola. The semi-major axis, represented by the distance OP in Figure 1, is therefore a measure of the curvature of the learning curve. The vertex, P, is that point at which all of the animals are eliminating one error per trial.

From Figure 1 we can readily see that, as the length of OP increases, the learning curve becomes flatter until it approaches a straight line. As the curve approaches a straight line OP approaches infinity. When the curve becomes a straight line we no longer have learning. The animal makes the same number of errors in every trial. He eliminates none of them.

On the other hand as *OP* approaches zero we find that the curvature becomes greater until it approaches an abrupt angle such as that made by the asymptotes themselves. The animal makes no correct responses and then makes all correct responses. There is sudden learning, an immediate change from "not known" to "known."

This is the situation which has been defined as "insight." This type of learning is characterized by the suddenness of its appearance, the absence of gradual change, a sudden change from random ineffectual behavior to immediate effectual behavior.

From the first papers in this series we have shown that the value of *OP* may be obtained from the expression

$$OP = \sqrt{2BC} \tag{4}$$

in which B and C are the constants of the learning curve found from fitting actual experimental data. From (4) the value of *OP* will be zero if either B or C are zero. In the case that B is zero (2) is of the form

$$u' = A. \tag{5}$$

When C is zero (2) becomes

$$u' = A + B. \tag{6}$$

Both (5) and (6) are straight lines parallel to the axis on which trials are measured.

In terms of the constants of (1) the length of *OP* is expressed by the relationship

$$OP = \frac{1}{K}\sqrt{\frac{2m}{a}}. \tag{7}$$

When we have no learning *OP* approaches infinity, k is zero or m is infinite. The animal has no ability to learn the problem on hand, or the problem is too difficult.

When we have "insight", *OP* approaches zero, k approaches infinity, or m approaches zero. In curves which show "insight" the learning ability of the animal for the particular situation is very great or the problem is very easy. Thus, our formulation lines up with that given by Thurstone.

Conclusions:

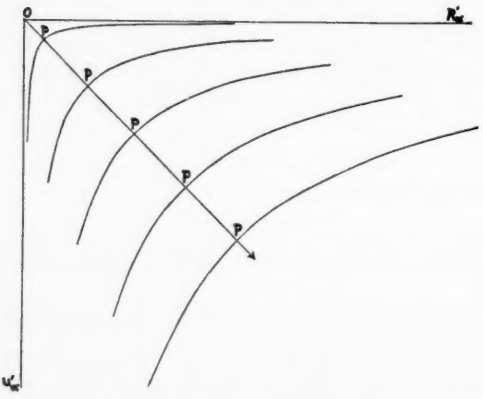
From the learning curve developed by L. L. Thurstone we have been able to derive a theoretical quantitative explanation of that type of learning called "insight."

1. The semi-major axis gives a mathematical expression for the limits of learning.

2. The limits of the semi-major axis are zero and infinity.

   a. When the semi-major axis approaches infinity there is no learning and the curve is a straight line. Either the animal has no ability to learn the problem set to him or the problem is too hard.

   b. When the semi-major axis approaches zero, learning is immediate, sudden, producing the situation called "insight." Either the learning ability of the animal for the situation is very great or the problem itself is very easy, which is the definition formerly made by Thurstone.



Trials are represented in the horizontal direction, accumulated errors in the vertical direction. OR'_a is the horizontal asymptote of the curve, Ou'_a is the vertical asymptote. O is the intersection of the asymptotes. P is the vertex of the curve and OP is the length of the semi-major axis.

REFERENCES

THURSTONE, L. L., "The Learning Function," *Jour. of Gen. Psychol.*, 1930, **3**, pp. 469-493.

THURSTONE, L. L., "The Error Function in Maze Learning," *Jour. of Gen. Psychol.*, 1933, **9**, pp. 288-301.

WILEY, L. E. AND WILEY, A. M., "Studies in the Learning Function. I. An Empirical Test of Thurstone's Theoretical Learning Curve," *Psychometrika*, 1937, **2**, pp. 1-19.

WILEY, L. E. AND WILEY, A. M., "Studies in the Learning Function. II. Critical Values of the Learning Curve", *Psychometrika*, June, 1937, **2**.

PSYCHOMETRIKA—VOL. 2, NO. 3
SEPTEMBER, 1937

SIMPLIFIED FORMULAS FOR ITEM SELECTION AND CONSTRUCTION

DOROTHY C. ADKINS
The University of Chicago

and

HERBERT A. TOOPS
The Ohio State University

The formula for the Pearson correlation coefficient of a dichotomous variable with a multiple-categoried variable is simplified for computational purposes by effecting in the multiple-categoried variable two types of arbitrary distributions: (1) rectangular and (2) proportional to binomial expansion coefficients. The formulas which result are convenient for the selection of test items and are applicable to the objective estimation of the comparative merits of the alternatives in multiple-choice test items. It is shown that the authoritative answer should have a high positive criterion coefficient, while the omissions and several wrong-answer alternatives should each have low (algebraic) negative criterion coefficients.

Where the intercorrelations of items are relatively low, the item-criterion correlation coefficient is an index for their selection and elimination which compares favorably with more complicated techniques. As shown below, such an index may be adapted not only to the task of determining the relative worth of items but also to that of objectively estimating the comparative merits of alternatives in the construction of multiple-choice test questions. From the standpoint of computation, any simplification of formulas facilitating this two-fold purpose will be of practical value. By using arbitrary forms of distribution of the criterion variable — a process which has but slight *relative* effect on the criterion coefficients of competing items— we have derived several very simple formulas for the correlation of a dichotomously-scored item and a multiple-categoried criterion. In our opinion, corrections for coarse grouping will be unnecessary.

The gross-score formula for the Pearson correlation coefficient may be written

$$r_{XY} = \frac{N\Sigma XY - \Sigma X \Sigma Y}{\sqrt{N\Sigma X^2 - (\Sigma X)^2}\,\sqrt{N\Sigma Y^2 - (\Sigma Y)^2}}. \qquad (1)$$

If X be an item scored dichotomously (right = 1; not right = 0) and administered with no time limit, certain simplifications in Formula (1) result, viz.,

$$\Sigma X = \Sigma X^2 = R, \tag{2}$$

$$\Sigma XY = \Sigma Y_R, \tag{3}$$

where R is the number of persons answering Item X correctly and ΣY_R is the sum of the criterion scores of those R persons.

Substituting (2) and (3) in (1), we have:

$$r_{XY} = \frac{N\Sigma Y_R - R\Sigma Y}{\sqrt{R(N-R)}\ \sqrt{N\Sigma Y^2 - (\Sigma Y)^2}}. \tag{4}$$

Certain restrictions imposed on the Y-distribution will simplify (4).* Since these restrictions depend on the type of distribution considered desirable, rectangular and approximately normal distributions are discussed separately below.

Type A. Rectangular Distribution of Criterion Scores.

Consider a rectangular criterion distribution of k categories, each of N/k scores.

The origin of the coded Y-scores is set at the middle of the range, with coded scores symmetric about zero; for example: $\cdots -2, -1, 0, 1, 2 \cdots$ when k is odd, or $\cdots -5, -3, -1, 1, 3, 5 \cdots$ when k is even. In either case, the term ΣY becomes zero, thus reducing (4) to

$$r_{XY} = \frac{N\Sigma Y_R}{\sqrt{R(N-R)}\ \sqrt{N\Sigma Y^2}}. \tag{5}$$

For a given study, the *relative* merit of items can be determined by

$$r_{XY}\sqrt{\frac{\Sigma Y^2}{N}} = \frac{\Sigma Y_R}{\sqrt{R(N-R)}} = \frac{\Sigma Y_R}{\sqrt{RW}}, \tag{6}$$

where $W = N - R$, and ΣY^2 and N are constants.

If the actual r's are desired, a further simplification of (5) results from a consideration of the term $\sqrt{\frac{N}{\Sigma Y^2}}$, which, when multiplied by the right-hand member of (6), gives r.

Where k is odd, it may be shown that

*With the appropriate assumptions, the modifications to be presented are equally applicable to the biserial correlation coefficient, as well as to other simple item indices, and they are adapted to either an external or an internal criterion.

$$\sqrt{\frac{N}{\Sigma Y^2}} = \sqrt{\frac{12}{k^2 - 1}} \tag{7}$$

and, where k is even, that

$$\sqrt{\frac{N}{\Sigma Y^2}} = \sqrt{\frac{3}{k^2 - 1}}. \tag{8}$$

Restricting k to 10 or less,* we may tabulate the corresponding values of k and $\sqrt{\frac{N}{\Sigma Y^2}}$ as follows:

TABLE 1

k	2	3	4	5	6	7	8	9	10
$\sqrt{\frac{N}{\Sigma Y^2}}$	1	$\sqrt{\frac{3}{2}}$	$\sqrt{\frac{1}{5}}$	$\sqrt{\frac{1}{2}}$	$\sqrt{\frac{3}{35}}$	$\frac{1}{2}$	$\sqrt{\frac{1}{21}}$	$\sqrt{\frac{3}{20}}$	$\sqrt{\frac{1}{33}}$

The most useful simplifications of (5) result when k is 2, 5 or 7. Where $k = 2$ and $\Sigma Y^2 = N$,

$$r_{XY} = \frac{\Sigma Y_R}{\sqrt{RW}}. \tag{9}$$

Where $k = 5$ and $\Sigma Y^2 = 2N$,

$$r_{XY} = \frac{\Sigma Y_R}{\sqrt{2RW}}. \tag{10}$$

Where $k = 7$ and $\Sigma Y^2 = 4N$,

$$r_{X1} = \frac{\Sigma Y_R}{2\sqrt{RW}}. \tag{11}$$

The size of N may be set arbitrarily by two considerations: (a) N must be an integral multiple of k and (b) it must be large enough to render sampling errors negligible. Values of N which fulfill these requirements are 1000 when k is 2 or 5 and 700 when k is 7. The resulting denominator terms may be tabled as a series of simple reciprocals, $\frac{1}{\sqrt{RW}}$ or $\frac{1}{\sqrt{2RW}}$ or $\frac{1}{2\sqrt{RW}}$, with R as the argument.

*Such a restriction is of value regardless of the use of automatic equipment, such as the Hollerith tabulating machinery, to which the recommended formulas are readily adapted.

Of the above three formulas, (10), employing five criterion categories, possibly best serves the practical and theoretical requirements.

Type B. Criterion Scores Distributed According to Binomial Expansion Coefficients.

We may take frequencies proportional to the binomial expansion coefficients, thus securing a distribution approaching normality, as, for example:

Y	—2	—1	0	1	2
Relative frequency	1	4	6	4	1

The Y-scores being thus coded symmetrically about zero, it is obvious that ΣY is again zero and that (5) and (6) apply for this case as well as for Type A. These formulas require that the sum of the relative frequencies, 2^{k-1}, be commensurable into N.* With the frequencies as assumed, there is no available formula for ΣY^2 nor for $\sqrt{\frac{N}{\Sigma Y^2}}$. However, the latter term may be computed directly and tabulated; thus:

TABLE 2

k	2	3	4	5	6	7	8	9	10
$\sqrt{\frac{N}{\Sigma Y^2}}$	1	$\sqrt{2}$	$\sqrt{\frac{1}{3}}$	1	$\sqrt{\frac{1}{5}}$	$\sqrt{\frac{2}{3}}$	$\sqrt{\frac{1}{7}}$	$\sqrt{\frac{1}{2}}$	$\frac{1}{3}$

The case $k = 2$ is, of course, equivalent to the two-categoried rectangular distribution. When $k = 5$, as in the distribution illustrated, (5) again reduces to its simplest form, (9).

Hence, summing the evidence to date, for convenience in solving correlation coefficients of a dichotomous variable with a k-categoried variable, the following distributions in the multiple-categoried variable are recommended:

(a) A rectangular distribution with $k = 2$, N even, and symmetrically coded criterion scores, where (9) applies. The resulting index probably is not so sensitive to item quality as that obtained for larger values of k.

*If N is too large, a random selection of cases may be discarded; and, if N is too small, possibly a few cases may be duplicated. One may also consider the possibility of altering the size of k.

(b) A rectangular distribution with $k = 5$, N a multiple of 5, and symmetrically coded criterion scores, where (10) applies (although $k = 7$, with N a multiple of 7, yields a computationally simpler formula).

(c) A distribution with relative frequencies 1, 4, 6, 4, 1, with $k = 5$, N a multiple of 16, and symmetrically coded criterion scores, where (9) again applies.

If, however, only an index proportional to r is desired, (6) may be used with any distribution of symmetrically coded criterion scores and an N which satisfies the requirements for the selected type of distribution.

The Evaluation of Item Alternatives.

There remains to be pointed out the adaptation of such formulas to the problem of evaluating alternatives for a multiple-choice item. The correlation of an item with a criterion is the correlation between the criterion and the *answer regarded as right by authoritative opinion.* If each alternative in turn be considered as "right", the same formulas are applicable to the computation of alternative-criterion coefficients.

As an example, assume that the criterion is coded 2, 1, 0, —1, —2; that the criterion distribution is rectangular; and that, for each criterion group, the frequency of choice of a given alternative is denoted by *a, b, c, d,* and *e,* respectively. Then the terms for the solution of a simplified correlation formula will be clear from the following schematic arrangement:

TABLE 3

(1)	(2)	(3)
Y	Frequency of Choice	$Y \cdot$ fr.
+2	a	$+2a$
+1	b	$+ b$
0	c	0
—1	d	$-d$
—2	e	$-2e$
Sum	C	ΣY_c

Where X_c refers to any alternative chosen, C, analogous to the former R for the "right" answer, is the number of persons making the choice in question and is obtained by summating the frequencies in

Column 2. $\sum Y_C$ is the sum of the products in Column 3. Replacing R by C and substituting the value of $\sum Y_C$ for $\sum Y_R$ in (10), the formula appropriate to the assumed distribution, we obtain as the formula for the alternative-criterion coefficient

$$r_{X_C Y} = \frac{\sum Y_C}{\sqrt{2C(N-C)}} = \frac{2a + b - d - 2e}{\sqrt{2C(N-C)}}. \qquad (12)$$

Formula (12) makes feasible the solution of the criterion coefficients for each of the responses obtained from a preliminary administration of items in the completion form.

In a trial of this technique, items were first administered to an experimental group by the completion method; as an example:

Misdirection : misdirect : : pessimism : _____

For this particular item, Table 4 gives the seventeen different responses and their respective frequencies, as obtained from 95 subjects. Formula (12) yields the values listed in the right-most column. The authoritative answer (No. 17) has a positive criterion coefficient, .16, which, however, is not so high as is desirable. Among the answers with positive coefficients is No. 7, "pessimise", which may be regarded as a misspelling of "pessimize." If the former version (No.

TABLE 4. Criterion Correlation Coefficient Solving-Table for the 17 Offered Answer Alternatives

No.	Alternatives Offered	Frequency of Choices of Criterion Group					Total C	$2a + b - d - 2e$	$r_{X_C Y}$
		2	1	0	−1	−2			
1	(Item Omitted)	3	1	1	3	2	10	0	.000
2	pessimistic	6	5	9	7	7	34	−4	−.062
3	pessify	2			2		4	+2	+.074
4	pessimist	1	6	6	3	6	22	−7	−.124
5	pessimity	1					1	+2	+.146
6	pess--------	1			1		2	+1	+.052
7	pessimise		2				2	+2	+.104
8	pessimism		1				1	+1	+.073
9	depress		1				1	+1	+.073
10	pessimify		1				1	+1	+.073
11	pessimitate		1				1	+1	+.073
12	pessimitus			1			1	0	.000
13	pessistic				1		1	−1	−.073
14	pessinate					1	1	−2	−.146
15	pessimatic					1	1	−2	−.146
16	pessimious					1	1	−2	−.146
17	PESSIMIZE	5	1	2	2	1	11	+7	+.163
	Total	19	19	19	19	19	95		
		a	b	c	d	e			

7) were omitted from the alternatives of a new multiple-choice form, it is to be expected that the persons of the experimental group who wrote "pessimise" would choose "pessimize." When the several *a*, *b*, *c*, *d*, and *e* frequencies for the two choices, 7 and 17, are combined, the predicted coefficient for "pessimize" in the new form becomes .19. It may also be argued that the persons making any of the remaining choices with positive coefficients would be more likely to choose "pessimize" than any of the alternatives with negative coefficients. Hence, if the alternatives with positive coefficients are omitted in the multiple-choice form, the *predicted* upper limit of the criterion coefficient for the authoritative answer (No. 17) is that obtained by combining the frequencies of the alternatives having positive coefficients (Nos. 3, 5, 6, 7, 8, 9, 10, 11) with the corresponding ones for "pessimize" (No. 17). Since this procedure yields an estimate of .31,* the actual coefficient to be realized from a reconstructed multiple-choice item will probably fall between .16 and .31.

Answers to be selected as "wrong" alternatives for the multiple-choice item are those with the highest negative coefficients. Since Table 4 contains four answers with coefficients of —.12 or less (algebraical) such a recasting of the illustrated item seems promising. If *N* be small, the sampling errors of the choice coefficients may be reduced by concentrating the choices on a few alternatives arbitrarily formulated by the test constructor.

Formula (12) also provides a convenient means of comparing one's predictions with the results actually obtained from a trial of the chosen alternatives in multiple-choice form. Further revisions can be made to eliminate those alternatives which violate this principle of item construction, viz., that the authoritative answer should have a high positive criterion coefficient, while the several wrong-answer alternatives and omissions should each have low (algebraic) negative criterion coefficients.

*This prediction assumes that the implied result actually will occur. It will be noted, also, that one may allow two different answers as correct and, by combining frequencies, determine the predictive value of allowing either, but not both, when checked by any examinee. Multiple testee responses, however, should be treated as omissions or resolved into a unitary score.

PSYCHOMETRIKA—VOL. 2, NO. 3
SEPTEMBER, 1937

THE DETERMINATION OF THE FACTOR LOADINGS OF A GIVEN TEST FROM THE KNOWN FACTOR LOADINGS OF OTHER TESTS

PAUL S. DWYER
University of Michigan

A technique is indicated by which approximations to the factor loadings of a new test may be obtained if factor loadings of a given group of tests and the correlations of the new test with the other tests are known. The technique is applicable to any orthogonal system and is especially adapted to cases in which $\Sigma a_{ji}\, a_{jk} = 0$ when $i \neq k$. Application is also made to the simultaneous determination of the factor weights of a group of tests in which no additional common factor is present. The technique is useful in adding tests to a completed factorial solution and in using factorial solutions involving errors to give results which are approximately correct.

It happens not infrequently that, after a set of intercorrelations has been subjected to a multiple factor analysis, additional correlations with other tests are available. One naturally wishes to enlarge his analysis to include this new material. The technique explained below is devoted to the development of a means of incorporating the new results without the disheartening necessity of repeating the whole factorial solution.

Suppose that the intercorrelations of tests 1, 2, $\cdots$ j, $\cdots$, r are subjected to a multiple factor analysis which results in k common factors. The resulting weights of each of the r tests, and the communality of each, are indicated in Table I,

TABLE I

Test	a_{j1}	a_{j2}	$\cdots$	a_{jk}	h_j^2
1	a_{11}	a_{12}	$\cdots$	a_{1k}	h_1^2
2	a_{21}	a_{22}	$\cdots$	a_{2k}	h_2^2
3	a_{31}	a_{32}	$\cdots$	a_{3k}	h_3^2
..	$\cdots$	$\cdots$	$\cdots$	$\cdots$	..
j	a_{j1}	a_{j2}	$\cdots$	a_{jk}	h_j^2
..	$\cdots$	$\cdots$	$\cdots$	$\cdots$	..
r	a_{r1}	a_{r2}	$\cdots$	a_{rk}	h_r^2

where the k reference vectors constitute any orthogonal system. Let t be a test whose correlations with the r tests are $r_{1t}, r_{2t}, \cdots, r_{rt}$ respectively.

To find the unknown weightings $a_{t1}, a_{t2}, \cdots a_{tk}$.

These weightings a_{ti} must approximately satisfy the equations

$$a_{t1}a_{11} + a_{t2}a_{12} + \cdots + a_{ti}a_{1i} + \cdots + a_{tk}a_{1k} = r_{1t}.$$
$$a_{t1}a_{21} + a_{t2}a_{22} + \cdots + a_{ti}a_{2i} + \cdots + a_{tk}a_{2k} = r_{2t}.$$
$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$
$$a_{t1}a_{j1} + a_{t2}a_{j2} + \cdots + a_{ti}a_{ji} + \cdots + a_{tk}a_{jk} = r_{jt}.$$
$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$
$$a_{t1}a_{r1} + a_{t2}a_{r2} + \cdots + a_{ti}a_{ri} + \cdots + a_{tk}a_{rk} = r_{rt}.$$

Applying the theory of least squares we get the normal equations

$$a_{t1}\Sigma a_{j1}^2 + a_{t2}\Sigma a_{j1}a_{j2} + \cdots + a_{ti}\Sigma a_{j1}a_{ji} + \cdots + a_{tk}\Sigma a_{j1}a_{jk} = \Sigma a_{j1}r_{jt},$$
$$a_{t1}\Sigma a_{j1}a_{j2} + a_{t2}\Sigma a_{j2}^2 + \cdots + a_{ti}\Sigma a_{j2}a_{ji} + \cdots + a_{tk}\Sigma a_{j2}a_{jk} = \Sigma a_{j2}r_{jt},$$
$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$
$$a_{t1}\Sigma a_{j1}a_{jk} + a_{t2}\Sigma a_{j2}a_{jk} + \cdots + a_{ti}\Sigma a_{j1}a_{jk} + \cdots + a_{tk}\Sigma a_{jk}^2 = \Sigma a_{jk}r_{jt},$$

whose solution gives the desired $a_{t1}, a_{t2}, \cdots a_{ti}, \cdots a_{tk}$.

The communality is $h_t^2 = a_{t1}^2 + a_{t2}^2 \cdots + a_{tk}^2$.

As an example consider a seven variable set of correlations used by Thurstone (**1**, page 129) together with the correlations of these seven variables with an eighth variable. The intercorrelations are presented in Table II.

TABLE II

	1	2	3	4	5	6	7	8
1	...	.58	—.28	.01	.36	.38	.61	.15
2	.58	...	—.28	.10	.52	.44	.78	.10
3	—.28	—.28	...	.56	.14	—.56	—.42	—.70
4	.01	.10	.56	...	.40	—.34	—.03	—.65
5	.36	.52	.14	.40	...	.08	.46	—.30
6	.38	.44	—.56	—.34	.08	...	.54	.50
7	.61	.78	—.42	—.03	.46	.54	...	.25
8	.15	.10	—.70	—.65	—.30	.50	.25	...

The principal axes solution is given by Thurstone (**1**, page 128) and is repeated in Table III with the undetermined test 8 and the values of r_{8j} indicated.

TABLE III

Test	$r_{j\Lambda_1}$	$r_{j\Lambda_2}$	$r_{j\Lambda_3}$	h^2	r_{8j}
1	+.659828	+.120945	.00	.45	.15
2	+.830332	+.265611	.00	.76	.10
3	—.541290	+.637969	.00	.70	—.70
4	—.126124	+.770774	.00	.61	—.65
5	+.437356	+.590526	.00	.54	—.30
6	+.637638	—.336776	.00	.52	.50
7	+.904489	+.109084	.00	.83	.25
8	$r_{8\Lambda_1}$	$r_{8\Lambda_2}$	.00	h_8^2	

Thurstone shows $\sum_{j=1}^{7} r_j^2{}_{\Lambda_1} = 2.849689$, $\sum_{j=1}^{7} r_j^2{}_{\Lambda_2} = 1.560312$ and it is easily shown in addition that

$$\sum_{j=1}^{7} r_{j\Lambda_1} r_{j\Lambda_2} = .000003, \quad \sum_{j=1}^{7} r_{j\Lambda_1} r_{8j} = 1.056625, \quad \sum_{j=1}^{7} r_{j\Lambda_2} r_{8j} = -1.221153,$$

so that the normal equations are

$$2.849689\, r_{8\Lambda_1} + .000003\, r_{8\Lambda_2} = 1.056625,$$

$$.000003\, r_{8\Lambda_1} + 1.560312\, r_{8\Lambda_2} = -1.221153,$$

and approximately

$$r_{8\Lambda_1} = \frac{1.056625}{2.849689} = .370786$$

$$r_{8\Lambda_2} = \frac{-1.221153}{1.560312} = -.782634$$

with $h_8^2 = .75$.

The solution of the normal equations is very simple when the non-diagonal coefficients in the normal equations are zero. This situation is attained in the illustration above and it appears (**2**, pages 425-426) in all cases in which a principal component solution is used. It is frequently approximately attained, when k is large, when other orthogonal axes, centroid for example, are used. Thus the non-diagonal terms of Table 7 (**1**, page 131) are small when compared with the diagonal entries. Approximate values of weights for an additional test t could be obtained by treating each non-diagonal entry as 0.

As a parenthetical remark we note that in case the non-diagonal entries are relatively small, they may be placed equal to zero. The diag-

onal entries then give first approximations to roots of the characteristic equation. The actual two decimal place values of the roots of the characteristic equation of Table 7 (**1**, page 131) and the approximations as determined by inspection are

TABLE IV

actual values	approximations
—5.02	—5.01
—1.18	—1.18
— .44	— .43
— .32	— .34

Suppose that a system of three orthogonal reference vectors is used to summarize the correlations of Table I. Thurstone has used such a system (**1**, page 124). This table, augmented to include the indicated weights of test 8 and the correlation of test 8 with the other tests is given in Table V.

TABLE V

Tests	a_{j1}	a_{j2}	a_{j3}	h_j^2	r_{8j}
1	+.5	—.2	+.4	.45	.15
2	+.6	—.2	+.6	.76	.10
3	—.6	+.5	+.3	.70	—.70
4	—.3	+.4	+.6	.61	—.65
5	+.2	+.1	+.7	.54	—.30
6	+.6	—.4	0	.52	.50
7	+.7	—.3	+.5	.83	.25
8	a_{81}	a_{82}	a_{83}	h_8^2	

From the principal axes solution it is expected that $h_8^2 = .75$. It is desired to find a_{81}, a_{82}, a_{83}. The normal equations become

$$1.95\, a_{81} - 1.07\, a_{82} + .69\, a_{83} = 1.165,$$

$$-1.07\, a_{81} + .75\, a_{82} + .11\, a_{83} = -.965,$$

$$.69\, a_{81} + .11\, a_{82} + 1.71\, a_{83} = -.565,$$

where the coefficients on the left have been previously found by Thurstone (**1**, page 125) in computing the characteristic equation. The solution of these equations is the easily verified.

$$a_{81} = .5,\ a_{82} = -.5,\ a_{83} = -.5 \text{ with}$$

$$a_{81}^2 + a_{82}^2 + a_{83}^2 = .75 \text{ as expected.}$$

A slight variation of the method makes possible the simultaneous computations of weightings on a number of tests. Denote the weightings on test t by $a_{t1}, \cdots, a_{tk}$ and the correlation of test t with the other tests by $r_{t1}, r_{t2}, \cdots r_{tr}$.

Let

$$R_{it} = \sum_{j=1}^{r} a_{ji}\, r_{tj} \qquad \text{and} \qquad \sum_{j=1}^{r} a_{i_1 j}\, a_{i_2 j} = A_{i_1 i_2}\,;$$

the normal equations are

$$\begin{aligned}
&A_{11}\, a_{t1} + A_{12}\, a_{t2} \cdots + A_{1j}\, a_{tj} + \cdots + A_{1k}\, a_{tk} = R_{1t}\,,\\
&A_{21}\, a_{t1} + A_{22}\, a_{t2} \cdots + A_{2j}\, A_{tj} + \cdots + A_{2k}\, a_{tk} = R_{2t}\,,\\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\\
&A_{k1}\, a_{t1} + A_{k2}\, a_{t2} \cdots + A_{kj}\, a_{tj} + \cdots + A_{kk}\, a_{tk} = R_{kt}\,.
\end{aligned}$$

If D is the determinant of the coefficients and D_{ij} is the cofactor of A_{ij}, then

$$a_{tj} = \frac{D_{1j}}{D} R_{1t} + \frac{D_{2j}}{D} R_{2t} + \cdots + \frac{D_{kj}}{D} R_{kt}\,.$$

The determinantal coefficients are the same no matter what the value of t. After they are computed it is only necessary to insert the R_{it} to find the different weightings. This method is discussed in more detail in an article which will be published soon **(3)**.

If a principal axes solution is found and another test is added to the solution by this method, the result will not give a principal axes solution of the new set of tests since the principal axes description of each test varies with the introduction or deletion of a test.

The general method outlined is also applicable when an error is found in one of the entries of the correlation matrix. For example, suppose that a mistake in sign was made in the value of r_{ij} which was recorded as +.55 rather than —.55. A mistake of this kind is apt to be discovered as the centroid solution advances since this residual may not approach zero as do the others.

In lieu of a complete repetition of the analysis, the following method is indicated. Take the result of the analysis which used the erroneous r_{ij} and cross out the weights for test i and test j. The weights on the following tests give the desired correlations with all the tests except test i and test j. Provide the weights for test i and test j by the method explained above using the correct correlations. The resulting centroid solution does not appear to give the same re-

sult as a centroid solution using the correct intercorrelations, but, when rotated to reveal simple structure, the results are in approximate agreement if the residuals of test i and j are of the same order as the other residuals.

REFERENCES

1. THURSTONE, L. L., *The Vectors of Mind.* University of Chicago Press, 1935.
2. HOTELLING, H., "Analysis of a Complex of Statistical Variables into Principal Components," *Journal of Educational Psychology,* Sept.-Oct., 1933, **24**.
3. DWYER, P. S., "The Simultaneous Computation of Groups of Regression Equations and Associated Multiple Correlation Coefficients,"*Annals of Mathematical Statistics,* December, 1937.

PSYCHOMETRIKA—VOL. 2, NO. 3
SEPTEMBER, 1937

A COMPARATIVE STUDY OF SCALES CONSTRUCTED BY THREE PSYCHOPHYSICAL METHODS*

MILTON A. SAFFIR
Bureau of Child Study, Board of Education, Chicago, Illinois

A comparison is made between the scales constructed by the Method of Paired Comparison, Rank Order, and the Method of Successive Intervals. Application of the three psychophysical methods to handwriting specimens and to nationality preferences results in mutually linear scales. Choice of scaling methods becomes, then, a matter of practical convenience rather than of relative validity.

The object of this study is to compare, on the basis of empirical data, the scales constructed through the use of the Method of Paired Comparison, the Rank Order Method, and the Method of Successive Intervals. All of these are developments of the traditional psychophysical methods which have made it feasible to construct scales for the quantitative study of psychological values that can not be objectively measured, such as social attitudes, nationality preferences, handwriting excellence, etc.

Since the psychological continua represented by attitude, preference, and judgment scales can not be directly measured, the validation of the psychophysical methods used in constructing the scales has depended upon a measure of the internal consistency in the method employed. With the Method of Paired Comparison, for example, the discrepancies between experimentally observed proportions and those calculated through the use of the scale values is such a measure of internal consistency. If we can allocate 25 stimuli along a linear scale in such a way that the 25 points account for some 300 experimental observations, within the limits of chance error, then we have grounds for using that scale to represent the psychological process responsible for the data.

We can further study the validity of a psychophysical method by comparing a scale constructed according to it with one constructed through the use of a different psychophysical method. While this does not constitute a final proof that either scale is valid for the purpose

*The writer is very much indebted to Professor L. L. Thurstone for the suggestion of this problem and for supervision in carrying out the study.

for which we propose to use it, yet if the two scales were comparable we could use both methods with increased confidence. If on the other hand the two scales were not comparable, we would find it necessary to determine which method was most valid for the description of the discriminatory process in question.

Miss Hevner (**1**) has carried through such a study of the Methods of Paired Comparison, Order of Merit (Rank Order Method) and Equal Appearing Intervals. She found that the scale values obtained through the use of the Method of Paired Comparison when plotted against those calculated from data from the Order of Merit Method, gave a linear plot, indicating that they were comparable. The scale values obtained by the Method of Equal Appearing Intervals when compared with those from the other two methods, gave plots which were not linear, indicating that the scales were not comparable. She concluded that scales constructed by the Paired Comparison Method or the Order of Merit Method were more dependable than those constructed by the Method of Equal Appearing Intervals, because the former corroborated each other, because in them the internal consistency could be demonstrated, because the latter was subject to error through the "end effect", and because the latter does not involve a test of internal consistency.

It should be pointed out that the agreement of the Methods of Paired Comparison and Order of Merit in Miss Hevner's study is not a check on the validity of the psychophysical methodology involved, since the same psychophysical process, Case V of Thurstone's Law of Comparative Judgment, was applied to the data obtained by both methods. All that was demonstrated was that the data obtained from the subjects' ranking the stimuli were comparable with the data obtained from the subjects' comparing each stimulus with every other one. Even if the logic used in deriving Thurstone's Law were so faulty as to invalidate its use in deriving a scale, Miss Hevner's results would still show the same close agreement between the two methods of gathering data which she employed. Therefore no judgment as to the validity of the scale values obtained from the data is justified by the agreement of scale values in her study.

The present investigation makes use of a new psychophysical process, the Method of Successive Intervals, which is entirely independent in its logic. This process, developed by Thurstone, is applicable to data gathered by the Method of Equal Appearing Intervals, takes account of the "end effect", and involves a test of internal consistency. It will be shown in this paper that this process is applicable

to data obtained by the Order of Merit Method, or, more properly, that the latter method can be treated as the special case of the Method of Successive Intervals in which the number of intervals is equal to the number of stimuli. We can compare the scales obtained from such an application with those obtained by Miss Hevner from the same data. Thus we can compare the scales which result from two different processes applied to a single set of data.

The present study, then, will deal with three methods of gathering data, and with two statistical procedures for treating the data. The comparisons will be between the scales resulting from the following four combinations:

(1) Paired Comparison data treated according to the Law of Comparative Judgment;
(2) Order of Merit data treated according to the Law of Comparative Judgment;
(3) Order of Merit data treated according to the Method of Successive Intervals;
(4) Equal Appearing (or Successive) Interval data treated according to the Method of Successive Intervals.

In addition to Miss Hevner's data dealing with judgments of excellence of handwriting specimens, a new set of data has been collected dealing with nationality and racial preferences. This makes it possible to compare the two psychophysical processes as applied to two different sorts of material, involving psychological activities that are quite distinct from each other.

Gathering The Data

1. The Stimuli and the Subjects.

In order to proceed with a comparison of the various psychophysical methods, it is necessary to use the same subjects and the same stimuli for each of the methods. In Miss Hevner's experiment the stimuli consisted of twenty specimens of handwriting, ranging from very poor to very excellent. Each of the 370 graduate and undergraduate students who served as subject was given some instruction as to the methods of judging the specimens, and then compared them according to the methodology of the three procedures — the Method of Paired Comparison, the Method of Equal Appearing Intervals, and the Order of Merit Method. These methods will be described below.

In the present experiment the stimuli consisted of the following 25 nationality and racial groups: American, Austrian, Belgian, Cana-

dian, Chinese, Englishman, Frenchman, German, Greek, Hindu, Hollander, Irishman, Italian, Japanese, Jew, Mexican, Negro, Norwegian, Pole, Russian, Scotchman, South American, Spaniard, Swede, and Turk. Nationality preference scales have been constructed by psychophysical methods a number of times, (**2**, **3**), and, as Thurstone has pointed out, the names of the above groups are valid psychological stimuli regardless of the questions that might be raised as to the genuineness of each as an ethnological unit.

In constructing a scale of nationality preferences, it is desirable to choose as homogeneous a population as possible. For this reason it was decided to use as subjects only native born, white Christians whose parents were also native born white Christians. Approximately 1000 sets of materials were distributed to University of Chicago undergraduate students in the elementary psychology and social science courses. Each set contained a paired comparison schedule and cards for the successive interval and rank order methods. The subjects were instructed to complete the materials for one method before beginning the second. They might sign their names or not, as they preferred. Of all the materials distributed, less than 400 sets were returned, and of these it was necessary to eliminate about two-thirds because of incompleteness or because the student was of Negro, Jewish, or foreign-born parentage. One hundred thirty-three returns were suitable according to the criteria that had been set up, and these were the only ones used in this study.

2. Method of Paired Comparison.

In the Method of Paired Comparison, which is the more complete form of the traditional Constant Method or Method of Right and Wrong Cases, the subjects compare each stimulus with every other one. For n stimuli there will be $\frac{n(n-1)}{2}$ comparisons for each subject to make. In the present study there were 300 pairs of nationalities on the paired comparison schedule.* In order that the data may be treated by some form of Thurstone's Law of Comparative Judgment, it is necessary that the subjects make a choice in each pair of stimuli, indicating that one is preferred to the other. Even where the two stimuli are psychologically very close together, so that they appear to the subject as just about equal, it is still necessary to make a choice—even if it appears to be more of a guess than a judgment. If the stimuli were actually

*Through a clerical error the combination Irishman-South American was omitted and Italian-South American appeared twice. The same number of subjects underlined each member of this pair at both places.

equal, and there were a large number of subjects who were compelled to make a choice, the laws of chance would operate, so that stimulus A would be preferred to stimulus B about as often as the reverse.

When all the schedules had been returned, the data were tabulated by counting the number of times each member of the 300 pairs of nationalities was underlined. Thus for the first combination, Japanese-Austrian, it was found that 16 subjects had underlined Japanese, 115 had underlined Austrian, and 2 had underlined neither. From this tabulation it was a simple step to prepare Table I which gives the proportion of the subjects which preferred each nationality to each one of the others.

3. Method of Successive Intervals.

The essential procedure in the Method of Successive Intervals is for the subject to sort the stimuli into a series of piles or groups, representing a succession of quantitative or qualitative values. This method is a more general form of the Method of Equal Appearing Intervals, in which the subjects follow the above sorting procedure with the additional condition that the piles must be so spaced as to form apparently equal steps or intervals. Data obtained through the more limited method can, of course, be treated by the successive interval procedure.

It is much easier for a subject to follow a sorting procedure if he has a fairly large number of specimens to sort, than if he has only two or three specimens to place in each pile. Hence in both Miss Hevner's and the present study, a good many additional stimuli were mixed with those used in the paired comparison part of the experiment. The subject was presented with all the stimuli in the same way, though the scale values were to be calculated for only the twenty or twenty-five that were used in the other methods.

Miss Hevner presented her subjects with 72 specimens of handwriting, instructing them to sort the samples into eleven piles in such a way that the difference in excellence represented by the interval between any two successive piles were the same. For each of the specimens she counted the frequency with which it was placed in each of the eleven piles. It is from these data that the scale values were recalculated in the present study.

The scaling procedure used by Miss Hevner for these data was quite simple. From the frequency distribution for each specimen a cumulative frequency curve was plotted, and the median, the point at which the curve crossed the 50% line, was read directly from the graph. These scale values are subject to the various criticisms listed

in Miss Hevner's conclusions which are quoted at the beginning of this paper.

To secure the successive interval data for the nationality preference study, the names of the twenty-five races and nationalities used in the paired comparison schedule were printed on separate cards, and in order to facilitate the sorting procedure for the subjects, there were added cards with the names of 35 additional groups. These materials were enclosed in a heavy manila envelope which was distributed to the subjects along with the paired comparison schedules. Enclosed in the envelope was the instruction sheet which asked the subjects to sort the nationalities into ten piles, placing in pile 1 those with which he would most prefer to associate, in pile 2 those which were a little less preferable, etc. To secure rank order data, the subjects were further instructed to number the nationalities in each pile in the order of preference.

The successive interval data was tabulated by counting the number of times each nationality was placed in each of the ten piles. Table II contains the frequency distribution for each of the twenty-five nationalities.

4. Rank Order Method.

The procedure in the Order of Merit or Rank Order Method is the simplest of all so far as the instructions to the subject are concerned. He is presented with a list of stimuli which he is told to put in rank order according to a given criterion. In Miss Hevner's study the subjects were told to rank the handwriting specimens in the order of their general excellence. As described above, instructions for ranking were included with the successive interval materials, in the nationality preference study. So many of the subjects, however, overlooked this part of the procedure, that as a result no rank order data are available for the nationality preference materials.

The tabulation of the rank order data depends on the statistical process which is to be applied. In Miss Hevner's study the Law of Comparative Judgment was used, so that it was necessary to construct a table of paired comparisons — that is, to determine the number of times each stimulus had been ranked more favorably than each of the other stimuli. This sort of a tabulation from rank order data is very laborious. If a subject rated the specimens $a, b, c, d, \cdots$ it was tabulated as a preferred to b, a preferred to c, a preferred to d, b preferred to c, b preferred to d, c preferred to d, etc. This sort of tabulation for the twenty stimuli had to be made separately for the rankings of each of the 370 subjects. From this point on Miss Hevner's

procedure was identical with her procedure in treating the data obtained by the Paired Comparison Method.

In a paper (**5**) published subsequent to Miss Hevner's study, Thurstone has developed a statistical technique for converting the raw data gathered by the rank order method into the paired comparison table necessary for the application of some form of the Law of Comparative Judgment. This method is much easier than the laborious counting procedure employed by Miss Hevner, and gives results that are very close approximations to those which she obtained. Thurstone calculated a set of scale values from Miss Hevner's data, using his shorter method, and found them almost identical with her scale values.

In the present study it was proposed to treat the rank order data by some method other than the Law of Comparative Judgment, used in both Miss Hevner's and Thurstone's papers. This was done by applying the statistical procedure of the Method of Successive Intervals, which is made possible by treating each absolute rank as an interval, so that there are as many piles as there are stimuli. Thus in applying this method to the Hevner data, we treat the material as if there were twenty piles, and each subject placed a single specimen in each pile. If a subject ranked the stimuli *a*, *b*, *c*, *d*, ··· it was treated as a case in which specimen *a* was placed in pile 1, specimen *b* in pile 2, specimen *c* in pile 3, etc. In tabulating the data in such a fashion we are finding the number of times stimulus *a* was ranked first, second, third, etc., and similarly for stimuli *b*, *c*, and each of the others. Such a tabulation is, of course, much easier to make than Miss Hevner's tabulation through counting. It is the first step according to Thurstone's simplified procedure.

Statistical Treatment of the Data

1. The Law of Comparative Judgment.

The Law of Comparative Judgment was first described by Thurstone in 1927 (**4**). Five cases of the law were described, differing in the number of assumptions that are made. In both Miss Hevner's and the present study, Case V, the simplest form of the Law of Comparative Judgment was used. This form, in addition to the assumptions made by the other cases, assumes that the discriminal dispersions for all the stimuli used are equal. The statistical form of Case V is: $S_1 - S_2 = x_{12}\sqrt{2}\,\sigma$, in which,

S_1 and S_2 are the scale values of stimuli 1 and 2, respectively

x_{12} is the sigma value corresponding to the observed propor-

tion of judgments "Stimulus 1 is greater than stimulus 2"

σ is the stimulus dispersion, assumed to be equal for all stimuli.

In order to use all the data, the above formula was converted to

$$S_1 - S_2 = \frac{\Sigma(x_{1k} - x_{2k})}{n},$$

in which the subscript k denotes each stimulus taken in turn, n is the number of stimuli for which x_{1k} and x_{2k} were used, and the unit of measurement is $\sqrt{2}\,\sigma$. The scale separation between each successive pair of nationalities was calculated from this formula. The arbitrary origin of the scale was set at the lowest nationality, Turk, and from the scale separations the scale value of each nationality was determined. Column two of Table III contains the scale values obtained for each of the twenty-five nationalities.

Having obtained the scale values, it is next necessary to check the internal consistency — that is, the extent to which the scale values account for the complete table of proportions. This can be done very simply by using the scale values as starting points, and reversing the whole process to arrive at a set of calculated proportions which may be compared with the experimental proportions. Since we now know the scale values, we can calculate the value of x_{12} in the equation:

$$S_1 - S_2 = x_{12},$$

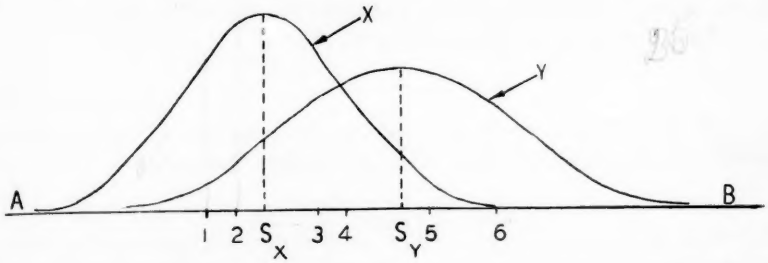
and with the use of the probability table, this value may be converted into the corresponding proportion. This is done for every pair of nationalities. The average discrepancy between the calculated and the original experimental proportions for the entire table was .031.

Miss Hevner used exactly the same procedure as has just been described in calculating the scale separations and final scale values for both her paired comparison and order of merit data. The scale values obtained by her are listed in columns two and three of Table IV. In checking the internal consistency as described above, Miss Hevner found an average discrepancy of .024 for the paired comparison data, and .012 for the order of merit data.

2. Method of Successive Intervals.

The statistical procedure for calculating the scale values by the Method of Successive Intervals has been described by Thurstone in his lectures, but has not yet been published.

The fundamental conditions in the method are as follows: The psychological scale is by definition such that the frequency distribution of the discriminal processes for each stimulus is normal. The discriminal dispersion (the standard deviation of each of the frequency distributions) varies for each stimulus. The scaling problem consists in determining the modal discriminal process for each stimulus. We have given a frequency distribution for each stimulus in terms of a given number of piles, the piles being successively, though not necessarily equally, spaced along the psychological scale.



The diagram above illustrates the problem. The line AB represents the psychological continuum. It is defined so that the curves $x, y, \cdots$ are normal. Points 1, 2, $\cdots$ represent the successive piles; the distance between them is unknown. The scaling problem consists of allocating $S_x, S_y, \cdots$ along AB. We are given the proportions for each stimulus corresponding to each one of the piles.

It is a simple step to determine the scale so that the distribution of x is normal; we simply read off the sigma values corresponding to the given cumulative proportions. If 25% of the subjects place stimulus x in pile 1, then point 1 will be located .67 sigma from S_x. In a similar manner we obtain the sigma values for the other piles.

Let us now follow a similar procedure with the cumulative proportions of distribution y. Since the psychological continuum is the same, the relative distances between the piles will remain the same, no matter what distribution is used. Two factors, aside from chance errors, will operate to change the sigma-values for points 1, 2, etc. Since the sigma-value for pile 1 obtained from distribution x gives its distance from S_x, and that obtained from distribution y gives its distance from S_y, the two values will differ by the distance $S_x - S_y$. This distance is a constant that is added to the sigma-value for each pile when it is obtained from distribution y rather than x.

The second factor which causes the sigma-value for the same point to vary with the distribution used, is the fact that the unit of measurement is the sigma of the particular distribution used; if σ_x is twice σ_y, then the sigma-value obtained for pile 1 from distribution x will be half that obtained from distribution y. If we plot the two sets of sigma values on a graph we will obtain a linear distribution, since the values are proportional to each other. The first factor mentioned above will be reflected in the fact that the curve will not pass through the origin; there will be a y-intercept equal to the difference $S_x - S_y$ in terms of σ_y as the unit. The second factor will be reflected in the slope of the line which will express the ratio of σ_x to σ_y.

The equation of a straight line is:

$$y = ax + b\,. \tag{1}$$

If we fit this curve to our data for distributions x and y by the method of averages, the line will pass through the points

$$\frac{\Sigma x_1}{n_1}, \frac{\Sigma y_1}{n_1}, \text{ and } \frac{\Sigma x_2}{n_2}, \frac{\Sigma y_2}{n_2}$$

so that the equation will take the form:

$$y = \frac{\frac{\Sigma y_1}{n_1} - \frac{\Sigma y_2}{n_2}}{\frac{\Sigma x_1}{n_1} - \frac{\Sigma x_2}{n_2}} x + b\,, \tag{2}$$

which reduces to

$$y = \frac{n_2 \Sigma y_1 - n_1 \Sigma y_2}{n_2 \Sigma x_1 - n_1 \Sigma x_2} x + b\,. \tag{3}$$

Since the slope of this curve is equal to the ratio $\frac{\sigma_x}{\sigma_y}$, we have

$$\frac{\sigma_x}{\sigma_y} = \frac{n_2 \Sigma y_1 - n_1 \Sigma y_2}{n_2 \Sigma x_1 - n_1 \Sigma x_2}\,. \tag{4}$$

Solving for σ_y, we have

$$\sigma_y = \frac{\sigma_x (n_2 \Sigma x_1 - n_1 \Sigma x_2)}{n_2 \Sigma y_1 - n_1 \Sigma y_2}\,, \tag{5}$$

from which we can determine the value of σ_y by making σ_x the unit of measurement, calling it 1.

We can also write (1) in the form

$$y = \frac{\sigma_x}{\sigma_y} x + b . \tag{6}$$

Solving for b we have

$$b = y - \frac{\sigma_x}{\sigma_y} x . \tag{7}$$

But since

$$b = \frac{S_x - S_y}{\sigma_y} \tag{8}$$

we have, passing the curve through $\frac{\Sigma x_1}{n_1}, \frac{\Sigma y_1}{n_1}$,

$$\frac{S_x - S_y}{\sigma_y} = \frac{\Sigma y_1}{n_1} - \frac{\sigma_x}{\sigma_y} \frac{\Sigma x_1}{n_1} . \tag{9}$$

This reduces to

$$S_y = S_x + \frac{\sigma_x \Sigma x_1 - \sigma_y \Sigma y_1}{n_1} . \tag{10}$$

Since our scale must have some arbitrary origin, let us set it at S_x by equating it to 0. We can now solve (10) for S_y. Knowing σ_y and S_y we can set up equations analogous to (5) and (10) for determining σ_z and S_z, and so on for each of the stimuli. We have thus a method for determining the scale value and stimulus dispersion for each of the stimuli used. The unit of measurement is σ_x, and the scale has its arbitrary origin at S_x.

There is, moreover, a check on the original assumption that the intervals between the piles could be so adjusted that all the distributions are normal. This will constitute a check of internal consistency. Since we know the sigma-value for each pile in each distribution, we can obtain the interval between any two piles for each distribution. We can reduce the intervals from all the distributions to a common unit, since the values of all the σ's are known in terms of σ_x. If our assumptions are correct, the intervals 1-2, 2-3, etc. will not differ markedly in the different distributions.

The scale values and stimulus dispersions obtained according to this method for the nationality preference data have been listed in columns 3 and 4 of Table III, those obtained from Miss Hevner's Order of Merit and Equal Appearing Intervals data appear in columns 4 to 7 of Table IV.

In each case in which the Method of Successive Intervals was used, the check on internal consistency mentioned above was made. There were few marked deviations in the several determinations of

the width of each of the piles; this bears out our fundamental assumption, that by allocating each pile to a definite position on the scale the distributions of all the stimuli become normal.

Comparison of the Results

Having obtained two sets of scale values for the nationality preference data (Table III) and four sets of scale values for the data from Miss Hevner's handwriting study (Table IV), the final step consists of comparing these sets graphically. Figure I provides a graphical comparison of the two scales obtained for the nationality preference study. Figures II-VII provide comparisons between the scales obtained for the handwriting study.

It is obvious from an inspection of these graphs, that the relationship between any two sets of scale values is a linear one. In several of the graphs there is very little scatter of the points; in others the scatter is somewhat wider, but in all of them the relationship is clearly a linear one. In several of the graphs the points at either extreme do not fall in a straight line, but it should be borne in mind that the end points are based on very few overlapping cases, and are thus unreliable.

Since the scatter of points is due to chance errors, it would be expected that there would be least scatter where there was least opportunity for chance errors. This would be expected for the graph of the two scales which involved only a single set of data. Inspection of Figure V indicates that the plot shows the least deviation or scatter of any of the graphs.

On the basis of the linearity of all the plots, we may conclude that all the methods employed in this study produce equally valid scales. Since the three different methods of gathering data—the Method of Paired Comparison, Order of Merit Method, and Method of Successive Intervals, — and the two different psychophysical techniques for scaling the raw data—the Law of Comparative Judgment and the Method of Successive Intervals — produce comparable scales, we may use any one with considerable confidence. In setting up a problem involving scaling, the choice of methods can be governed by matters of convenience, rather than by questions of relative validity.

It is of some practical interest to discover that the Order of Merit procedure in gathering data combined with the Method of Successive Intervals in treating the raw data, gives a scale comparable with any of the others. The material set-up as well as the labor on the part of

the subjects for a paired comparison study is great, and where the number of stimuli is large, is almost prohibitive. The Order of Merit Method where there are less than thirty stimuli, and the Method of Successive Intervals where there are more than thirty seem to be the most convenient methods of gathering data.

The use of the Law of Comparative Judgment — even when Case V, the simplest form, is employed — is more laborious than the technique of the Method of Successive Intervals. For rank order data, even with Thurstone's shortcut for converting the data into a paired comparison table, it is less laborious to employ the Method of Successive Intervals than the Law of Comparative Judgment.

REFERENCES

1. HEVNER, KATE, "Three Psychophysical Methods," *Journal of General Psychology*, 1930, **4**, pp. 191-212.
2. PETERSON, R. C. and THURSTONE, L. L. *Motion Pictures and Social Attitudes of Children*, New York: Macmillan, 1933.
3. THURSTONE, L. L., "An Experimental Study of Nationality Preferences," *Journal of General Psychology*, 1928, **1**, pp. 405-425.
4. THURSTONE, L. L., "A Law of Comparative Judgment," *Psychological Review*, 1927, **34**, pp. 273-286.
5. THURSTONE, L. L., "Rank Order as a Psychophysical Method," *Journal of Experimental Psychology*, 1931, **14**, pp. 187-201.

TABLE I

Proportion of Subjects which Preferred Nationality at Head of Each Column to Nationality in Each Row

	American	Austrian	Belgian	Canadian	Chinese	Englishman	Frenchman	German	Greek	Hindu	Hollander	Irishman	Italian	Japanese	Jew	Mexican	Negro	Norwegian	Pole	Russian	Scotchman	So. American	Spaniard	Swede	Turk
American		.030	.008	.060	.000	.106	.015	.053	.015	.023	.008	.091	.015	.008	.015	.015	.015	.023	.023	.008	.098	.015	.023	.038	.000
Austrian	.970		.556	.931	.068	.947	.679	.880	.105	.120	.636	.805	.242	.122	.203	.135	.060	.662	.241	.256	.842	.323	.271	.659	.000
Belgian	.992	.444		.947	.030	.962	.561	.765	.045	.061	.621	.692	.218	.030	.165	.045	.015	.511	.105	.211	.744	.226	.295	.561	.038
Canadian	.940	.069	.053		.008	.557	.218	.286	.008	.030	.098	.303	.053	.015	.030	.030	.015	.120	.053	.053	.250	.053	.053	.182	.008
Chinese	1.000	.932	.970	.992		1.000	.940	.970	.689	.422	.985	.962	.880	.568	.647	.583	.298	.977	.750	.827	.985	.824	.872	.924	.336
Englishman	.894	.053	.038	.443	.000		.090	.233	.008	.015	.030	.260	.038	.008	.015	.038	.000	.038	.015	.015	.135	.038	.023	.045	.000
Frenchman	.985	.321	.439	.782	.060	.910		.669	.053	.045	.394	.684	.113	.045	.114	.053	.045	.432	.128	.120	.729	.165	.165	.602	.030
German	.947	.120	.235	.714	.030	.767	.331		.023	.023	.242	.451	.120	.008	.038	.045	.000	.308	.053	.076	.485	.083	.121	.188	.000
Greek	.985	.895	.955	.992	.311	.992	.947	.977		.290	.947	.932	.856	.341	.556	.466	.189	.947	.687	.735	.962	.789	.805	.909	.120
Hindu	.977	.880	.939	.970	.578	.985	.955	.977	.710		.955	.947	.932	.585	.712	.586	.368	.977	.786	.856	.985	.856	.895	.910	.353
Hollander	.992	.364	.379	.902	.015	.970	.606	.758	.053	.045		.609	.136	.060	.128	.061	.023	.568	.090	.195	.788	.195	.233	.545	.015
Irishman	.909	.195	.308	.697	.038	.740	.316	.549	.068	.053	.391		.068	.061	.083	.045	.023	.295	.090	.120	.545		.120	.290	.030
Italian	.985	.758	.782	.947	.120	.962	.887	.880	.144	.068	.864	.932		.226	.323	.258	.075	.871	.402	.459	.931	.534	.580	.826	.030
Japanese	.992	.878	.970	.985	.432	.992	.955	.992	.659	.415	.940	.939	.774		.609	.515	.229	.947	.684	.714	.962	.833	.895	.917	.254
Jew	.985	.797	.835	.970	.353	.985	.886	.962	.444	.288	.872	.917	.677	.391		.432	.173	.932	.526	.591	.970	.684	.795	.880	.235
Mexican	.985	.865	.955	.970	.417	.962	.947	.955	.534	.414	.939	.955	.742	.485	.568		.260	.932	.684	.759	.970	.802	.835	.902	.188
Negro	.985	.940	.985	.985	.702	1.000	.955	1.000	.811	.632	.977	.977	.925	.771	.827	.740		.985	.818	.870	1.000	.893	.908	.947	.580
Norwegian	.977	.338	.489	.880	.023	.962	.568	.692	.053	.023	.432	.705	.129	.053	.068	.068	.015		.121	.135	.662	.203	.143	.538	.015
Pole	.977	.759	.895	.947	.250	.985	.872	.947	.313	.214	.910	.910	.598	.316	.474	.316	.182	.879		.629	.924	.639	.744	.887	.114
Russian	.992	.744	.789	.947	.173	.985	.880	.924	.265	.144	.805	.880	.541	.286	.409	.241	.130	.865	.371		.962	.504	.568	.850	.098
Scotchman	.902	.158	.256	.750	.015	.865	.271	.515	.038	.015	.212	.455	.069	.038	.030	.030	.000	.338	.076	.038		.076	.105	.256	.015
So.American	.985	.677	.774	.947	.176	.962	.835	.917	.211	.144	.805		.466	.167	.316	.198	.107	.797	.361	.496	.924		.466	.803	.098
Spaniard	.977	.729	.705	.947	.128	.977	.835	.879	.195	.105	.767	.880	.420	.105	.205	.165	.092	.857	.256	.432	.895	.534		.820	.061
Swede	.962	.341	.439	.818	.076	.955	.398	.812	.091	.090	.455	.710	.174	.083	.120	.098	.053	.462	.113	.150	.744	.197	.180		.053
Turk	1.000	1.000	.962	.992	.664	1.000	.970	1.000	.880	.647	.985	.970	.970	.746	.765	.812	.420	.985	.886	.902	.985	.902	.939	.947	

TABLE II

Frequency Distribution for each of Twenty-five Nationalities.
Method of Successive Intervals.

Nationality	1	2	3	4	5	6	7	8	9	10
American	127	4	1	1						
Austrian	12	21	30	27	29	7	5	1		1
Belgian	7	43	40	31	8	3	1			
Canadian	77	48	6	2						
Chinese		1	8	5	13	15	18	29	20	24
Englishman	91	38	4							
Frenchman	21	52	36	17	3	1		1		2
German	49	45	29	4	3	2	1			
Greek		5	10	12	19	26	22	19	15	5
Hindu		2	5	5	10	12	16	24	35	24
Hollander	10	45	45	20	8	2	2	1		
Irishman	46	42	29	10	2	2	1		1	
Italian	2	12	26	27	20	22	15	5	3	1
Japanese		1	8	6	14	18	19	24	20	23
Jew	6	12	14	15	11	19	14	15	10	17
Mexican		3	7	10	11	27	26	25	14	10
Negro		3	5	8	4	7	12	12	19	63
Norwegian	9	49	49	19	2	2	2		1	
Pole	3	7	13	27	20	22	18	9	4	10
Russian	2	19	16	17	29	19	14	8	8	1
Scotchman	51	49	25	6	2					
So. American	4	13	15	21	26	24	12	8	9	1
Spaniard	5	5	30	37	29	15	9	3		
Swede	18	48	37	11	11	1	5	1		1
Turk			3	1	3	9	17	24	48	28

TABLE III

Scale Values and Discriminal Dispersions Obtained through the Paired Comparison and Successive Intervals Methods. Nationality Preference Study.

Nationality	Scale Values Paired Comparison Method	Method of Successive Intervals	
		Scale Values	Discriminal Dispersions
American	4.1972		
Austrian	1.7348	2.6971	1.343
Belgian	1.9593	2.2500	.992
Canadian	3.0002	.2728	1.034
Chinese	.4125	5.6501	1.924
Englishman	3.2512	.0000	1.000
Frenchman	2.2020	1.7486	1.201
German	2.5918	1.0284	1.462
Greek	.7126	4.6351	1.609
Hindu	.3243	5.9533	1.965
Hollander	2.0815	2.1279	1.122
Irishman	2.5196	1.1637	1.502
Italian	1.2847	3.4891	1.602
Japanese	.5713	5.5448	1.917
Jew	.8821	4.3902	2.473
Mexican	.6489	5.0213	1.575
Negro	.0383	7.2720	3.219
Norwegian	2.1583	2.0439	1.064
Pole	1.0027	4.1406	1.823
Russian	1.2129	3.6581	1.896
Scotchman	2.5749	.9054	1.433
So. American	1.3202	3.7851	1.816
Spaniard	1.4223	3.2459	1.191
Swede	2.1023	1.9293	1.464
Turk	.0000	6.4682	1.580

TABLE IV

Scale Values and Discriminal Dispersions Obtained through Paired Comparison, Order of Merit, and Successive Intervals Methods. Handwriting Study.

Specimen	*Paired Comparison*	*Order of Merit* Law of Compar. Judgment	*Order of Merit Data* Method of Successive Intervals		*Successive Intervals*	
			S	σ	S	σ
B5	0.0000	.0000	0.0000	1.00	0.0000	1.00
G3	.4431	.4548	.5170	1.00	.4013	.90
H3	.4510	.6471	.7698	.81	.7009	.75
Y5	.4664	.8593	.8803	1.02	.6748	.99
F4	.5548	.6160	.7355	.85	.4574	.84
K2	.6679	.4336	.4677	1.03	.2614	.94
T4	1.4315	.8041	.8709	1.02	.4982	.94
F2	1.7219	1.9442	1.8339	.74	1.4672	.72
N2	1.8829	2.6589	2.5377	.90	1.9696	.83
B4	2.4637	2.8801	2.6603	.87	1.8902	.85
T3	3.2037	4.2595	3.8441	.82	2.8301	.77
Z3	3.3046	3.7929	3.5152	.81	2.6025	.82
J2	3.5206	4.2171	3.8605	.78	2.8045	.76
X3	4.3280	4.9039	4.4015	.79	3.0676	.77
R2	4.4369	5.4426	4.8556	.86	3.3676	.89
U3	5.5504	6.8529	6.0383	.62	4.0846	.75
A4	5.5302	6.7741	6.0072	.67	4.0490	.79
G4	5.8731	7.3983	6.4716	.83	4.3592	1.08
J3	7.4342	8.2750	7.2430	.87	4.8075	.81
D2	7.6887	8.5833	7.6938	1.01		

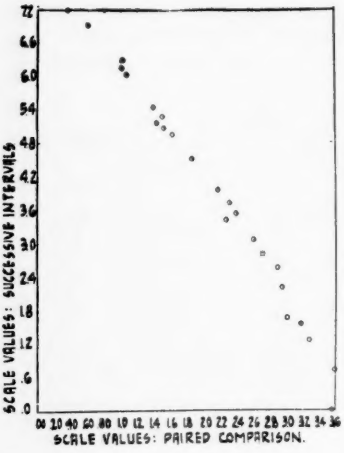


FIGURE I

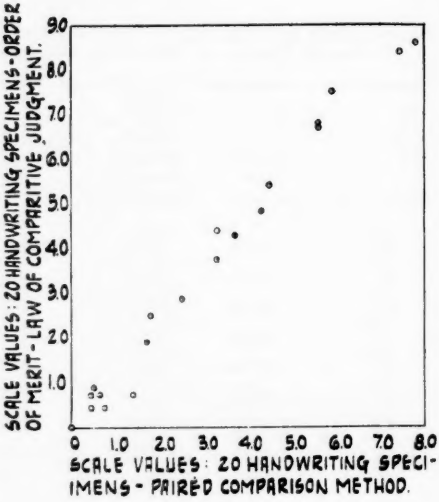


FIGURE II

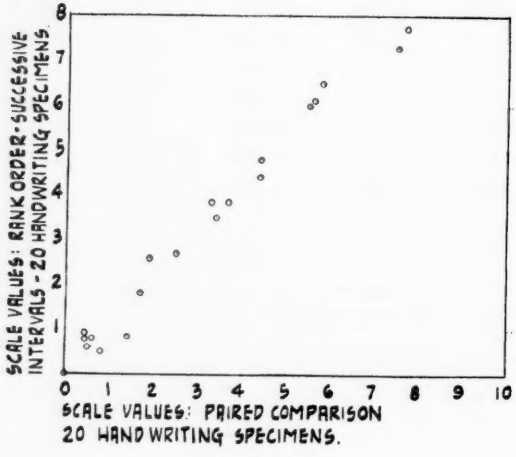


FIGURE III

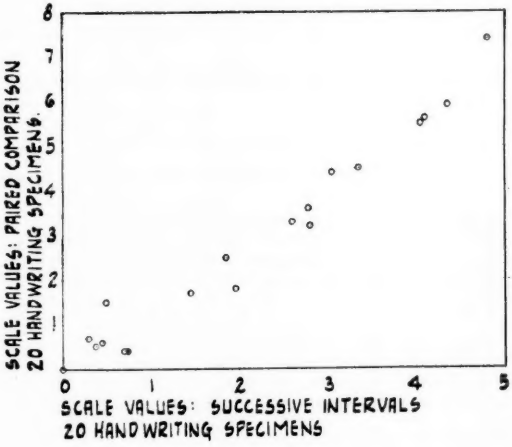


FIGURE IV

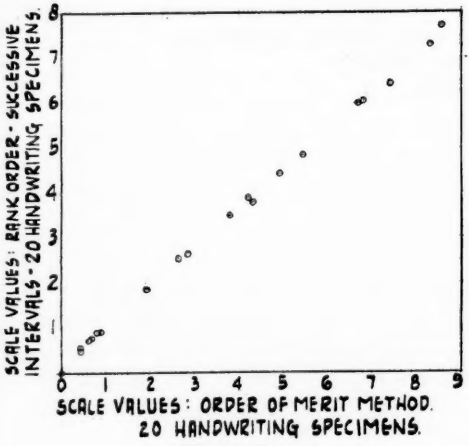


FIGURE V

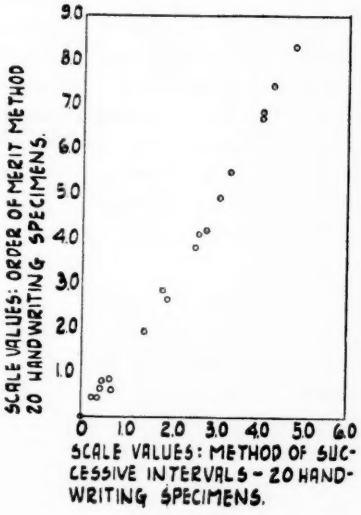


FIGURE VI

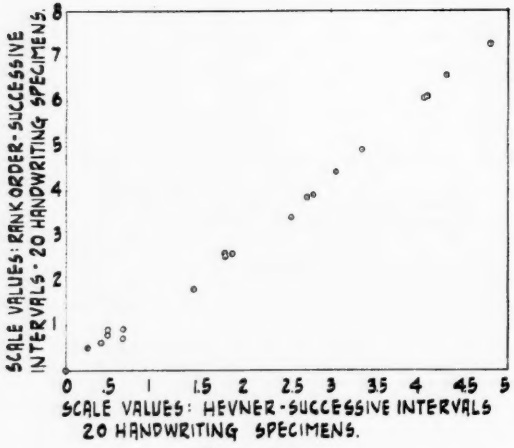

SCALE VALUES: RANK ORDER-SUCCESSIVE INTERVALS - 20 HANDWRITING SPECIMENS.
0
1
2
3
4
5
6
7
8
0
.5
1
1.5
2
2.5
3
3.5
4
4.5
5
SCALE VALUES: HEVNER-SUCCESSIVE INTERVALS
20 HANDWRITING SPECIMENS.

FIGURE VII

PSYCHOMETRIKA—VOL. 2, NO. 3
SEPTEMBER, 1937

MATHEMATICAL BIOPHYSICS OF CONDITIONING

N. RASHEVSKY
The University of Chicago

It is shown that from a somewhat more precise form of the fundamental postulates, used in a previous paper as a starting point for a system of mathematical biophysics of psychological phenomena, a mechanism can be derived, which represents the important features of conditioning.

In a previous paper, published in this journal and hereinafter referred to as I[1], we have outlined a general program for the development of mathematical biophysics in psychology. In a subsequent paper, referred to as II,[2] some particular points of the first, concerning conditioned reflexes were developed more in details. However in both those papers, while it was shown, that phenomena of conditioning are to be considered only as a particularly complex case of ordinary physical hysteresis,[3] properties of the mechanism, responsible for conditioning were not derived from the fundamental postulates, but had to be introduced as special assumptions. In the present paper we propose to show that no such additional assumptions of a physical nature are necessary and that provided we make our fundamental postulates somewhat more precise, the phenomena of conditioning can be derived from those, plus an assumption of a purely hystological nature. We now restate first our fundamental assumptions.

I.

While a continuous excitation of a nerve-fiber by means of a constant current usually results in a release of a simple excitation impulse, the qualitative and quantitative nature of which is entirely independent of the stimulus, provided the latter is strong enough to exceed the threshold, the situation for physiological stimuli, such as pressure, light, etc. is indifferent. Studies by Adrian and others show that a continuous stimulus sets off a volley of nerve-impulses following each other at approximately equal intervals. The stronger the

[1] N. Rashevsky: *Psychometrika*, **1**, 1, 1936.
[2] N. Rashevsky: *Psychometrika*: **1**, 265, 1936.
[3] N. Rashevsky: *ZS. f. Phys.* **53**, 102, 1929; *Jl. Gen. Psych*, **5**, 207, 1931; **5**, 368, 1931.

stimulus the shorter those intervals, or the higher the "frequency" of the sequence of impulses. While each individual impulse may be entirely independent of the intensity of the stimulus, preserving thus the "all or none law", yet the phenomenon as a whole gives a graded response to a graded stimulus. For not too strong intensities of the external stimulus the frequency of the volleys is approximately proportional to the intensity of the stimulus. We shall introduce the concept of the intensity of excitation E of a fiber, defining E as a quantity proportional to the frequency ν of impulses and to the intensity I of each individual impulse, Thus

$$E = I\,\nu. \tag{1}$$

If the "all or none law" holds for all fibers, then I is independent of the intensity S of the external stimulus and is a constant, characteristic of the fiber. ν however is proportional to S, but since S must in general exceed a threshold h in order that excitation would be released at all, ν is zero for $S = h$. Hence

$$\nu = a(S - h) \tag{2}$$

where however the factor of proportionality a may vary from fiber to fiber. (1) and (2) give

$$E = a\,I(S - h). \tag{3}$$

From physiological considerations it follows that (2) can at best be only an approximation. The interval between two successive impulses cannot be smaller than the refractory time θ of the fiber. Hence

$$\nu < \frac{1}{\theta} \tag{4}$$

no matter how strong S. As S increases indefinitely, ν must asymptotically tend to the value $\frac{1}{\theta}$, while for small values of S equation (2) holds approximately. There is of course an infinite number of functions which satisfy this requirement. Inasmuch as we have no empirical data yet to guide our choice, we shall choose the simplest possible one, namely:

$$\nu = \frac{1}{\theta}\left[1 - e^{-a\theta(S-h)}\right]. \tag{5}$$

For small values of $(S - h)$ we have

$$e^{-a\theta(S-h)} = 1 - a\theta(S - h)$$

and therefore (5) reduces to (2).

Using instead of (1) and (2) now (1) and (5) we obtain

$$E = \frac{I}{\theta}\left[1 - e^{-a\theta(S-h)}\right]. \tag{6}$$

With increasing S, E tends to I/θ.

We shall now consider the mechanism of transmission of excitation or the mechanism of inhibition. We assume that at the end of every excitatory fiber an *excitatory factor* ε is produced, according to the equation

$$\frac{d\varepsilon}{dt} = AE - a\varepsilon \tag{7}$$

where A and a are positive constants. We do not give any particular physical interpretation to this factor ε. It may be a special substance secreted in the immediate neighborhood of the end of the axon. Or it may be some other physico-chemical quantity, which follows *approximately* equ. (7). The development of the consequences of (7) is quite independent of such assumptions. In the future, of course, a physical interpretation of all equations, introduced here formally, such as (6) or (7) is to be attempted. For an inhibitory fiber we assume similarly, that its end produces an inhibitory factor j, according to

$$\frac{dj}{dt} = BE - bj \tag{8}$$

where B and b are positive constants.

For constant E, ε and j increase according to

$$\varepsilon = \frac{A}{a}(1 - e^{-at}); \quad j = \frac{B}{b}(1 - e^{-bt})$$

(c.f. I) and tend asymptotically to $\frac{A}{a}$ and $\frac{B}{b}$ respectively. $\frac{A}{a}E$ represents the maximum value, that ε can reach for a continuous excitation of strength E.

From (7) and (8) it follows, that when for a long time the fibers are not excited at all, the values of ε and j are appreciably zero. For in case $E = 0$ we have

$$\frac{d\varepsilon}{dt} = -a\varepsilon \quad \text{and} \quad \frac{dj}{dt} = -bj$$

which gives upon integration

$$\varepsilon = \varepsilon_0 e^{-at} \quad ; \quad j = j_0 e^{-bt}$$

ε_0 and j_0 being some initial values. Both ε and j tend to zero in the absence of excitation of corresponding fibers. Since A, B and E are always positive, ε and j are also always positive.

We may consider a more general case, in which an excitatory as well as an inhibitory fiber both produce ε and j. Only an excitatory fiber always produces an excess of ε, an inhibitory fiber always an excess of j. For simplicity we shall consider here the above more restricted form of postulates.

Now we must specify the effect of ε and j on a neighboring, adjacent fiber, that lies in the vicinity of the end of the axon, which produces either ε or j. We shall postulate that whenever $\varepsilon > j$, $\varepsilon - j$ acts as an excitatory stimulus on any such adjacent fiber. That is if h_2 is the threshold of such an adjacent fiber, then if $\varepsilon - j > h_2$, this fiber becomes excited, with an intensity given by (3) or more generally by (6), in which $\varepsilon - j$ is substituted for S. For small $\varepsilon - j$ we thus have for the intensity E_2 of the adjacent fiber

$$E_2 = a_2 I_2 (\varepsilon - j - h_2). \tag{9}$$

When $\varepsilon - j < h_2$, and a fortiori when $\varepsilon < j$, no excitation occurs.

II.

Let us now consider the following structure, which is suggested by numerous neurological observations. An arrangement of neurones, forming a "closed circuit", as represented in Fig. 1, has frequently been observed and described. In this arrangement a neurone is stimulated by another and the latter in turn stimulates the first neurone. Various possible significances of such an arrangement have been discussed. Let us consider it from the point of view which interests us.

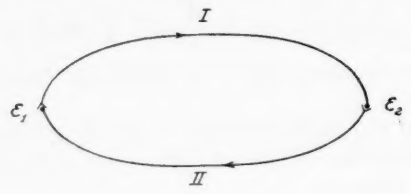


FIGURE 1

Let the threshold of I (Fig. 1) be h_1, that of II — h_2. In the absence of any external stimulation, none of the two neurones will be excited. Let however at the left end an amount of ε_1 be present, such

that $\varepsilon_1 > h_1$. This will produce a finite intensity of excitation E_1 in *I*, which in its turn will result in a production of ε_2 at the right end. If $\varepsilon_1 - h_1$ is sufficiently small, E_1 also will be sufficiently small and therefore ε_2 will be less than h_2. Hence *II* will not be excited. Under those conditions, when ε_1 again acquires the value zero, E_1 will become zero. Everything returns to its original state of non-excitation. Let however ε_1 exceed h_1 by such a large amount, that E_1 and therefore also ε_2 would become so large, that $\varepsilon_2 > h_2$. Then *II* will also become excited with an intensity E_2, which will result in the production of additional ε at the left end. This in its turn will result in an increase of E_1 and therefore in an increase of ε_2. The latter again increases E_2 and the process would thus tend automatically to infinity, if the linear relations between intensity of stimulation and intensity of excitation would hold exactly. As we have seen in section I, actually E_1 and E_2 tend to upper limits $\frac{I_1}{\theta_1}$ and $\frac{I_2}{\theta_2}$. Therefore the above described "self energizing" process will also actually stop, when E_1 and E_2 cannot increase any further, in spite of an increase of ε. But if now we bring the initial ε back to zero, it is possible that the additional ε produced by E_2 will be large enough to maintain E_1 excited and the system will remain in a continuous state of excitation even in the absence of an external stimulus.

These rather crude general considerations are borne out by mathematical analysis. To simplify the problem, we shall assume, that both *I* and *II* are very short and that the velocity of propagation of the excitation in both of them is very large, so that the time t_p which it takes for any variation of intensity of excitation at one end to reach the other is very small. In fact we consider it to be so small, that during this time neither ε_1 nor ε_2 can change appreciably.

We have

$$\left.\begin{aligned} E_1 &= \frac{I_1}{\theta_1}\left[1 - e^{-a_1\theta_1(\varepsilon_1 - h_1)}\right] \\ E_2 &= \frac{I_2}{\theta_2}\left[1 - e^{-a_2\theta_2(\varepsilon_2 - h_2)}\right] \end{aligned}\right\} \tag{10}$$

which gives

$$\left.\begin{aligned} \frac{d\varepsilon_1}{dt} &= \frac{AI_2}{\theta_2}\left[1 - e^{-a_2\theta_2(\varepsilon_2 - h_2)}\right] - a\varepsilon_1 \\ \frac{d\varepsilon_2}{dt} &= \frac{AI_1}{\theta_1}\left[1 - e^{-a_1\theta_1(\varepsilon_1 - h_1)}\right] - a\varepsilon_2. \end{aligned}\right\} \tag{11}$$

The analytic solution of the nonlinear system (11) is not known. We shall therefore investigate its properties by a graphical method.

$\frac{d\varepsilon_1}{dt} \gtreqless 0$ when

$$\frac{AI_2}{\theta_2}\left[1 - e^{-a\theta_2(\varepsilon_2 - h_2)}\right] - a\varepsilon_1 \gtreqless 0$$

or

$$\varepsilon_1 \lesseqgtr \frac{AI_2}{a\theta_2}\left[1 - e^{-a_2\theta_2(\varepsilon_2 - h_2)}\right]. \tag{12}$$

The equality sign in (12) gives the equation of a line, which is zero for $\varepsilon_2 = h_2$ and tends asymptotically to $\frac{AI_2}{a\theta_2}$ with increasing ε_2. It is represented by the full line in Fig. 2. For all points below that line $\frac{d\varepsilon_1}{dt} > 0$, for all points above it $\frac{d\varepsilon_1}{dt} < 0$.

Similarly, $\frac{d\varepsilon_2}{dt} \gtreqless 0$, when

$$\frac{AI_2}{\theta_1}\left[1 - e^{-a_1\theta_1(\varepsilon_1 - h_1)}\right] - a\varepsilon_2 \lesseqgtr 0$$

which may be written thus:

$$1 - e^{-a_1\theta_1(\varepsilon_1 - h_1)} \gtreqless \frac{a\theta_1}{AI_1}\varepsilon_2$$

or

$$e^{-a_1\theta_1(\varepsilon_1 - h_1)} \lesseqgtr 1 - \frac{a\theta_1\varepsilon_2}{AI_1} = \frac{AI_1 - a\theta_1\varepsilon_2}{AI_1} < 1.$$

Taking logarithms:

$$-a_1\theta_1(\varepsilon_1 - h_1) \lesseqgtr \log\frac{AI_1 - a\theta_1\varepsilon_2}{AI_1} < 0.$$

Hence

$$a_1\theta_1(\varepsilon_1 - h_1) \gtreqless \log\frac{AI_1}{AI_1 - a\theta_1\varepsilon_2} > 0$$

or finally

$$\varepsilon_1 \gtreqless h_1 + \frac{1}{a_1\theta_1}\log\frac{AI_1}{AI_1 - a\theta_1\varepsilon_2}. \tag{13}$$

The sign of equality in (13) represents a line, shown by the broken line in Fig. 2. For $\varepsilon_2 = 0$, $\varepsilon_1 = h_1 > 0$. As ε_2 increases, ε_1 increases

also. When $\varepsilon_2 = \frac{AI_1}{a\theta_1}$, the denominator of the log becomes zero, and $\varepsilon_1 = \infty$. For still larger values of ε_2, ε_1 has no real values.

For all points below that curve $\frac{d\varepsilon_2}{dt} < 0$; for all points above $\frac{d\varepsilon_2}{dt} > 0$.

If, as represented on Fig. 2 the two curves intersect at all for $\varepsilon_1 > 0$ and $\varepsilon_2 > 0$, then they intersect at two points G_1 and G_2. For values of ε_1 and ε_2 corresponding to these two points $d\varepsilon_1/dt = 0$ and

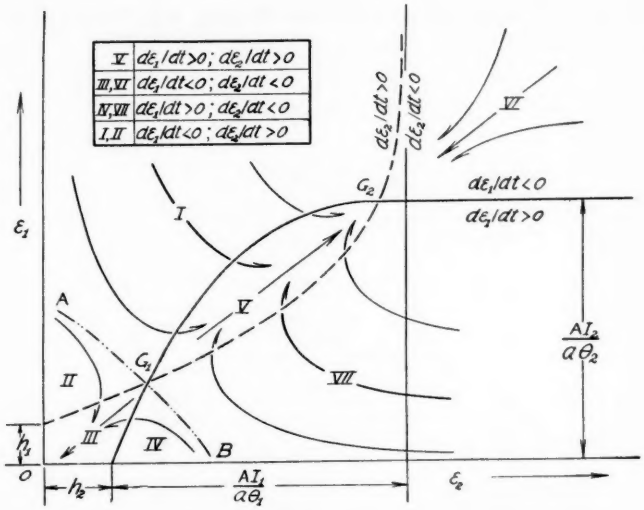


FIGURE 2

$d\varepsilon_2/dt = 0$. Hence the system does not change and is in equilibrium. It is however seen from inspection of Fig. 2 that while the configuration G_2 is stable, G_1 is unstable. Let the configurational point $(\varepsilon_1, \varepsilon_2)$ of the system be displaced from G_2 into region VI. Here as we see $d\varepsilon_1/dt < 0$ and $d\varepsilon_2/dt < 0$, therefore both ε_1 and ε_2 will decrease until they reach G_2 as shown by the arrows. If the point $(\varepsilon_1, \varepsilon_2)$ is in region V, then $d\varepsilon_1/dt > 0$ and $d\varepsilon_2/dt > 0$, the system moves again to G_2. In region I $d\varepsilon_1/dt < 0$, but $d\varepsilon_2/dt > 0$; the configurational point moves as indicated by the arrows until it comes into region V where it moves as we have seen to G_2. In region VII $d\varepsilon_1/dt > 0$, $d\varepsilon_2/dt < 0$ and we

have a similar situation. Fig. 2 indicates clearly, that while for any *small* displacement from G_2 the system returns to G_2 for any *small* displacement from G_1 it will move either to $\varepsilon_1 = \varepsilon_2 = 0$, or to G_2. The only exception is for displacements along the line AB, for which the system does return to G_1.

The above results can be demonstrated analytically by expanding the right-hand sides of (8) around G_1 and G_2, and keeping only the linear terms. We then obtain in the immediate vicinity of G_1 and G_2 for ε_1 and ε_2 a system of ordinary linear equations, the stability of whose solutions is determined and studied in the usual way. In this way it is also proven that the slope of the line AB at the point G_1 is equal to $-\sqrt{a_1 I_1 / a_2 I_2}$. This analytical method will be described elsewhere.

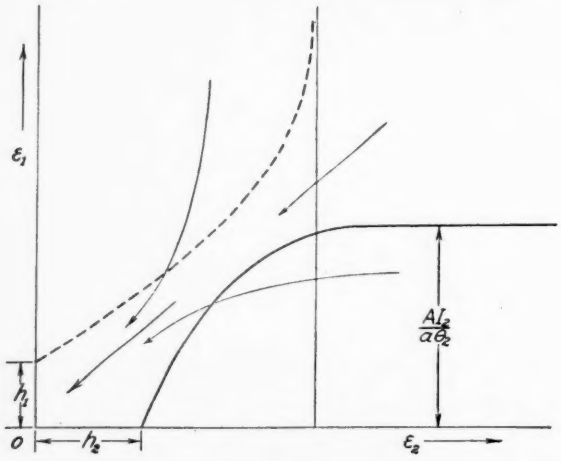


FIGURE 3

Thus the system considered has in this case in the absence of any external stimulation, two stable states of equilibrium. One corresponds to $\varepsilon_1 = \varepsilon_2 = 0$, the other to $\varepsilon_1 = \varepsilon_{0_1}$; $\varepsilon_2 = \varepsilon_{0_2}$. Then there is a line AB of unstable equilibrium, dividing the two states. As soon as by any external disturbance the system which was originally in a state $\varepsilon_1 = \varepsilon_2 = 0$ is brought into a state represented by a point to the right of AB, it "tips over" into G_2 $(\varepsilon_{0_1}, \varepsilon_{0_2})$ and remains there after the removal of the external disturbance.

If, as represented on Fig. (3) the full and broken lines do not

intersect at all, then in the absence of any external stimulation the only stable state is $\varepsilon_1 = \varepsilon_2 = 0$. Whether we shall have the case of Fig. (2) or that of Fig. (3) depends merely on the numerical values of the constants involved in our equations. The physical meaning of the case represented in Fig (3) is that the limiting value $\frac{I_1}{\theta_1}$ of E_1 is so small that even when it is reached, $\varepsilon_2 = \frac{A}{a}\frac{I_1}{\theta_1}$ is still less than h_2 and *II* therefore does not get excited, or that the limiting value $\frac{I_2}{\theta_2}$ of E_2 is too small to excite *I*.

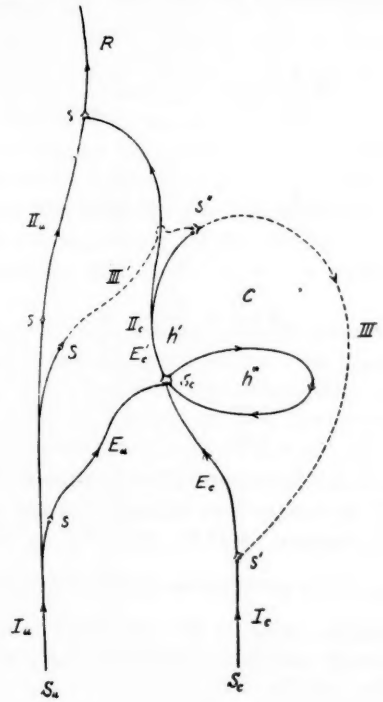


FIGURE 4

III.

Let us consider now a structure represented on Fig 4. Let the stimulus S_u produce an excitation in the nerve I_u with a threshold h_u, which carried over a number of synapses s finally results in some

reflex or reaction R. Let another series of nerves, I_c with threshold h_c, stimulated by a different stimulus S_c, lead to the synapse s_c at which it connects with a nerve II_c, which when excited also produces R through a final common path. Let this synapse s_c be excited also by a collateral of I_u and let it also connect to one end of a circuit, such as studied in the preceding section. Let that circuit be in an unexcited state and let the external excitation, which is necessary to bring it into an excited state, be $h^* \cdot h^*$ then represents the hysteresis threshold of this circuit. Let h' be the threshold of II_c and let ε_0 be the value of ε at the left end of the circuit, when it is in the stable excited state. Let furthermore the limiting vaues $\frac{I_u}{\theta_u}$ of E_u and $\frac{I_c}{\theta_c}$ of E_c be such that

$$P\frac{I_c}{\theta_c} < h' \quad ; \quad P = \frac{A}{a} \tag{14}$$

and

$$P\frac{I_c}{\theta_c} < h^* \quad ; \quad P\frac{I_u}{\theta_u} < h^* \tag{15}$$

but that

$$P\left(\frac{I_u}{\theta_u} + \frac{I_c}{\theta_c}\right) > h^* \tag{16}$$

and

$$P\frac{I_c}{\theta_c} + \varepsilon_0 > h'. \tag{17}$$

If a stimulus $S_u > h_u$ is applied, the reflex R is produced. However a stimulus S_c no matter how strong does not under those conditions produce R on account of (14), since PE_c is the maximum value that ε ever reaches for a continuous constant excitation E_c, and $\frac{I_c}{\theta_c}$ is the maximum possible value of E_c. If however both S_u and S_c are applied simultaneously and kept continuous for a sufficient time, then the amount ε and s_c will be

$$PE_u + PE_c = P(E_u + E_c).$$

On account of (16), this will become larger than h^*, if S_u and S_c are sufficiently strong, so that E_u and E_c are close enough to their limiting values I_u/θ_u and I_c/θ_c.

This will bring the circuit C into an excited state, in which in the absence of any external stimuli an amount ε_0 of ε will be produced

at the synapse s_c. When now again S_c is applied *alone* the total amount of ε at the synapse s_u will be for a continuous stimulation

$$P\frac{I_c}{\theta_c}+\varepsilon_0$$

and this is on account of (17) enough to excite II_c and therefore to produce a reflex R. Because of (15) application of S_u alone will not bring C into an excited state. The simultaneous application of S_u and S_c is essential.

We have here some of the principal features of Pavlov's conditioned reflex. In this simple scheme many fundamental details are however still missing.

We obtain some new features by complicating somewhat our scheme. If for instance on the efferent side of the circuit C the fiber II_c gives off a collateral, which through a synapse s'' excites an inhibitory fiber III, which ends at the synapse s', then after s_c has become conducting, a continuous stimulation of I_c produces a gradual increase of j at s' and thus inhibits the conditioned response. We have here the elements of internal inhibition of a conditioned reflex. If the fiber II_u sends off through a collateral an inhibitory fiber III' to the synapse s'', which inhibits the excitation of the inhibitory fiber III, then a simultaneous stimulation of I_u and I_c does not result in an inhibition of s'.

By integrating equ. (7) and (8), which govern the variation of ε and j at s_c and at other synapses for various types of stimulations a number of quantitative relations concerning the increase of the response R with the number of repetitions of simultaneous stimulations of I_u and I_c and other relations, may be derived along the lines indicated in I. The discussion of those relations must be reserved for a separate publication.